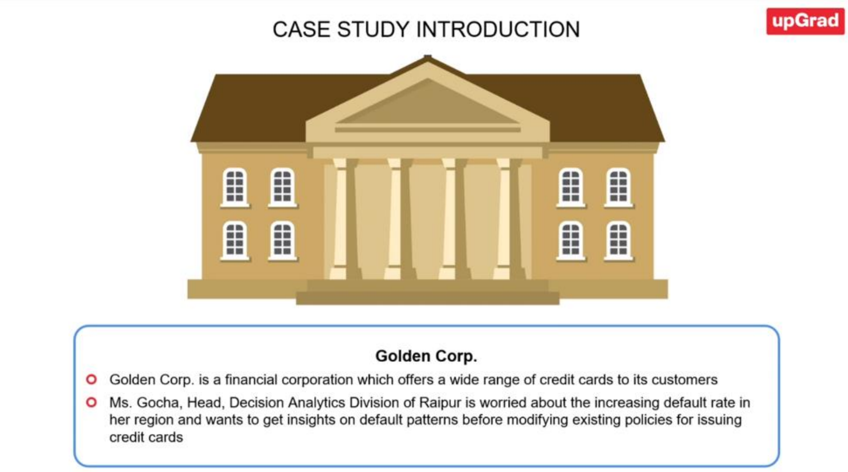
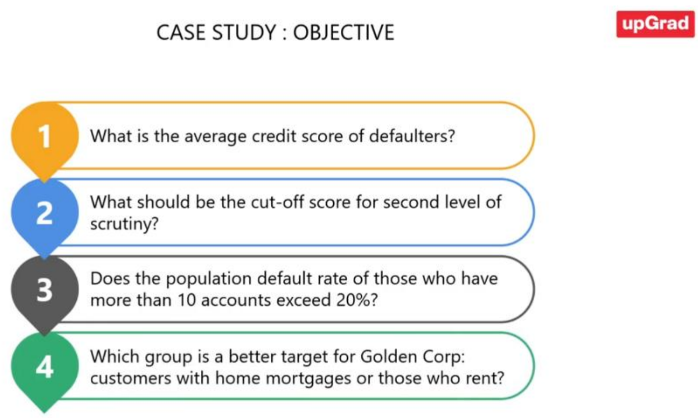
Transcription

Case Study: Analysis of Credit Defaulters



In this video, we will discuss application of interval estimation and hypothesis testing through a practical example. We will look at the case of Golden Corp, which is a financial institution offering wide range of credit cards.

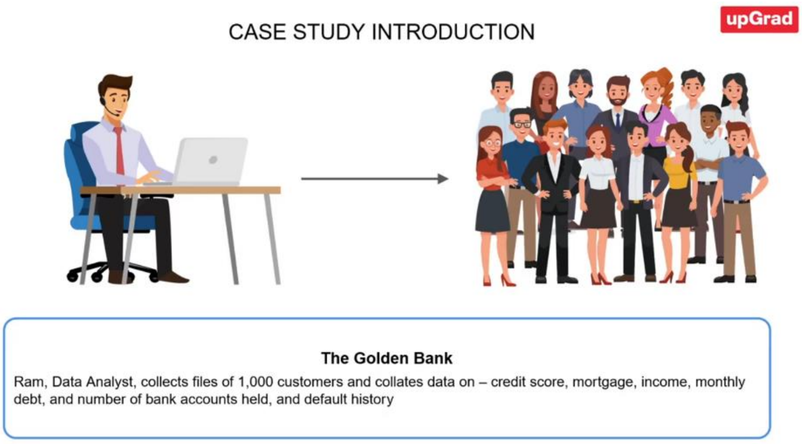
Within Golden Corp, we have Ms. Gocha who heads a decision analytics team in Raipur. She is worried about the default rate in her region and thinks that the existing policies of the company need to be modified. Before doing that, she wants to analyse the historical data and get some insight on default and spending patterns of credit card holders.



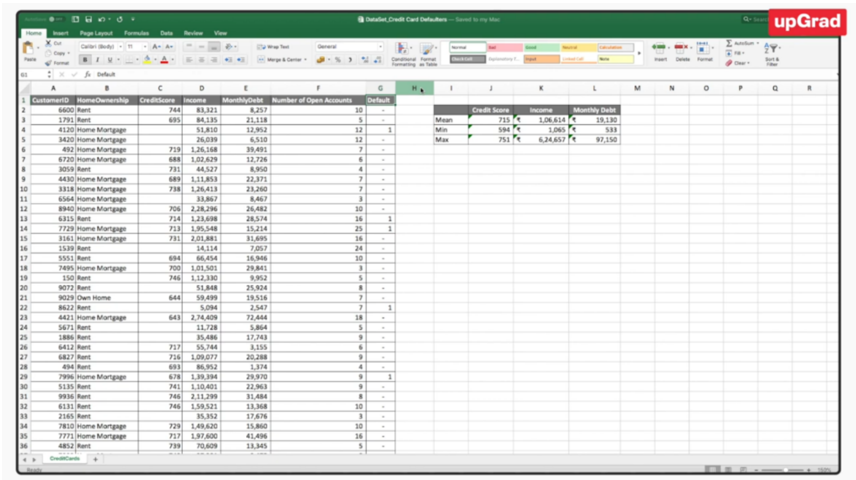
She goes on to discuss the project with her colleague Ram, and they decide on some of the more important objectives. The objectives are:

- One, what is the average credit score of defaulters?
- Two, what should be the cut-off score for a second scrutiny?
- Three, if the population default rate of those who have more than 10 bank accounts exceed 20%.

- And four, which group is a better target for Golden Corp, customers with home mortgages or customers who rent their houses?



We look at each of these objectives in greater detail. After discussing the objectives, Ram collects random files of 1000 credit card holders from the data team. He collates information on credit score, mortgages, income, monthly spending, number of bank accounts held and customers default history.

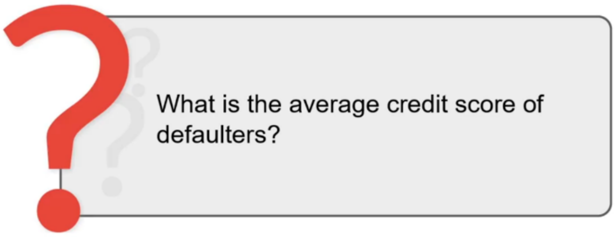


CustomerID	HomeOwnership	CreditScore	Income	MonthlyDebt	Number of Open Accounts	Default
6600	Rent	744	83,321	8,257	10	-
1791	Rent	695	84,135	21,118	5	-
4120	Home Mortgage		51,810	12,952	12	1
3420	Home Mortgage		26,039	6,510	12	-
492	Home Mortgage	719	1,26,168	39,491	7	-
6720	Home Mortgage	688	1,02,629	12,726	6	-
3059	Rent	731	44,527	8,950	4	-
4430	Home Mortgage	689	1,11,853	22,371	7	-
3318	Home Mortgage	738	1,26,413	23,260	7	-
6564	Home Mortgage		33,867	8,467	3	-
8940	Home Mortgage	706	2,28,296	26,482	10	-
6315	Rent	714	1,23,698	28,574	16	1
7729	Home Mortgage	713	1,95,548	15,214	25	1
3161	Home Mortgage	731	2,01,881	31,695	16	-
1539	Rent		14,114	7,057	24	-
5551	Rent	694	66,454	16,946	10	-
7495	Home Mortgage	700	1,01,501	29,841	3	-
150	Rent	746	1,12,330	9,952	5	-
9072	Rent		51,848	25,924	8	-
9029	Own Home	644	59,499	19,516	7	-
8622	Rent		5,094	2,547	7	1
4421	Home Mortgage	643	2,74,409	72,444	18	-
5671	Rent		11,728	5,864	5	-
1886	Rent		35,486	17,743	9	-
6412	Rent	717	55,744	3,155	6	-
6827	Rent	716	1,09,077	20,288	9	-
494	Rent	693	86,952	1,374	4	-
7996	Home Mortgage	678	1,39,394	29,970	9	1
5135	Rent	741	1,10,401	22,963	9	-
9936	Rent	746	2,11,299	31,484	8	-
6131	Rent	746	1,59,521	13,368	10	-
2165	Rent		35,352	17,676	3	-
7810	Home Mortgage	729	1,49,620	15,860	10	-
7771	Home Mortgage	717	1,97,600	41,496	16	-
4852	Rent	739	70,609	13,345	5	-

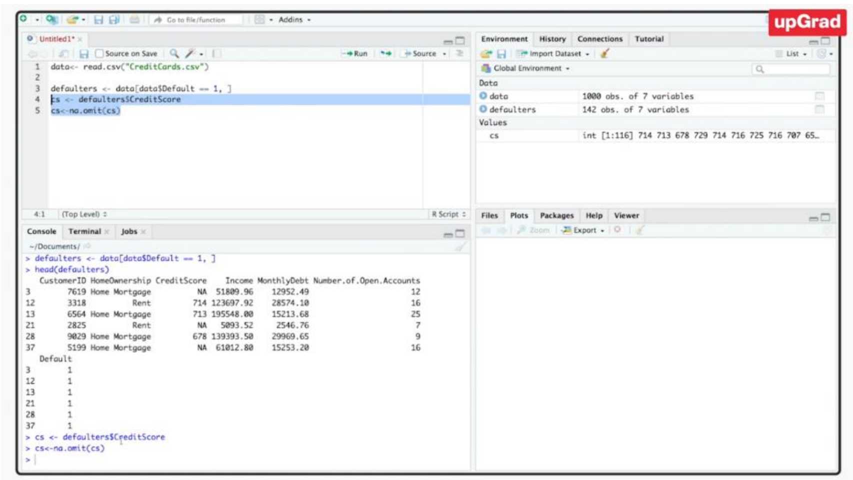
	Credit Score	Income	Monthly Debt
Mean	715	₹ 1,06,614	₹ 19,130
Min	594	₹ 1,065	₹ 533
Max	751	₹ 6,24,657	₹ 97,150

Let us take a look at the data. We have customer ID. We have home ownership, whether the customer owns a house, rents it or has a home loan. Then, we have credit score of customers, his or her monthly income, monthly debt or average credit card bill of the customer.

Then, there is number of open accounts, which is the total number of bank accounts held by the customer. Finally, we have default, which is one if the customer has defaulted on three or more consecutive bill payments and zero otherwise.



Coming back to the objectives, here is the first one, what is the average credit score of defaulters? How do we go about this? We have a sample data set, and based on the sample, we can estimate the 95% confidence interval for population mean of defaulters' credit score.



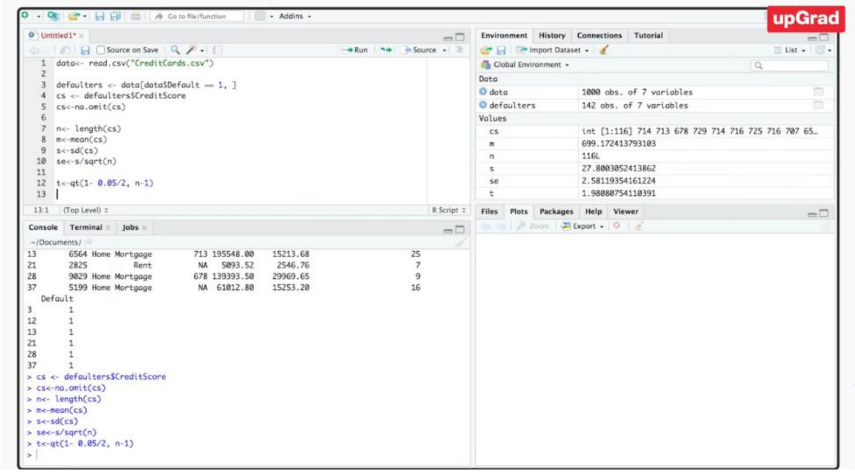
Let's look at that in R. So, we start with reading the data, we again use read.csv function here. The data is called credit card dot csv. Once we run it, we can do head data. We can also see the data here in this window. It has 1000 observation, which is on sample size.

Now, we want to estimate the confidence interval for credit score of defaulters. So, first we would need to subset the data. We only need the data of defaulters and defaulters would be all the data points or all the customers for whom default is equal to 1.

So, net subset R data, we have formed defaulters, which is data such that data dollar default equals to 1, and we can again do head defaulters. So, we have all the customers for whom default is 1.

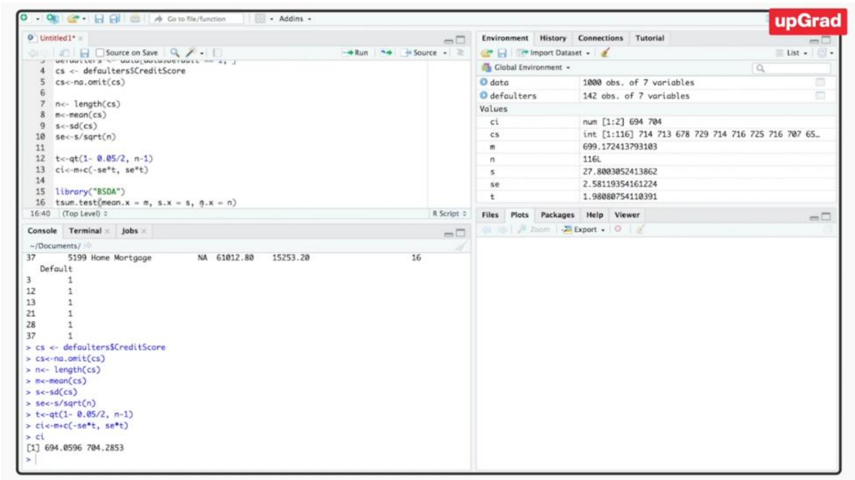
Next, we want to estimate confidence interval for credit score, and you can see there are NAs here. So, before estimating the confidence interval, we would need to get rid of these NAs.

Let us default cs as the credit score of defaulters or the series of these credit scores, and remember how we get rid of the NAs. We will use na.omit function. So, we will write na.omit cs. So, once we run it, we have a sample of defaulters of credit score.



Now, let us estimate the 95% confidence interval for mean credit score amongst defaulters. So, for sample size would be length of cs and the mean can be calculated using mean cs, which is defined as n here. We have s or standard deviation equal to sdcs, and then we compute the standard error, which is s divided by square root of n.

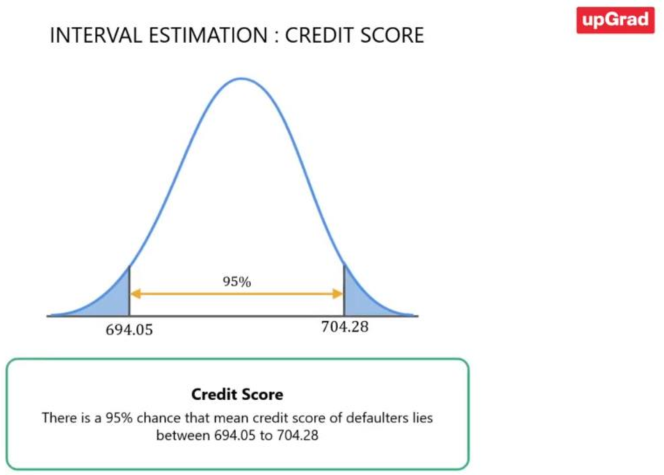
Now, as population standard deviation is not known. We will use T distribution and for 95% confidence interval, alpha would be 0.05, and alpha by 2 would be 0.025. So, the T value at 95% confidence interval can be calculated using QT function. As we want to estimate two-side confidence interval, we will use alpha by 2 and the degree of freedom for this T distribution would be n minus 1, which is the sample size minus one.



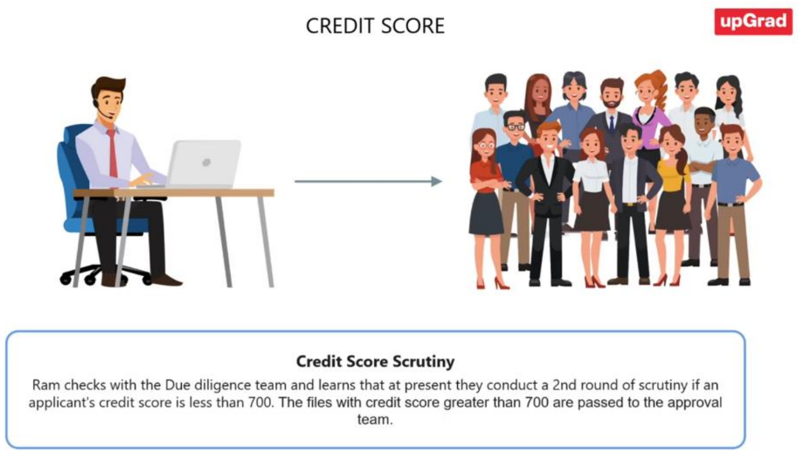
Now, we have T, we have standard error, and we have mean, so we can compute the confidence interval by adding and subtracting se into t from the mean. So, here is the confidence interval, which is 694.05 to 704.28.

Remember, we can also use t sum dot s function here. For that, we first need to call the library BSDA, and then we can call ts dot test function, mean dot x would be n, s dot x or standard deviation of sample would be s, and n dot x as the sample size.

If we run it, we have the confidence interval 694.05 to 704.28. So, based on this sample, we are 95% confident that the mean credit score of defaulters lie between 694 and 704.3.

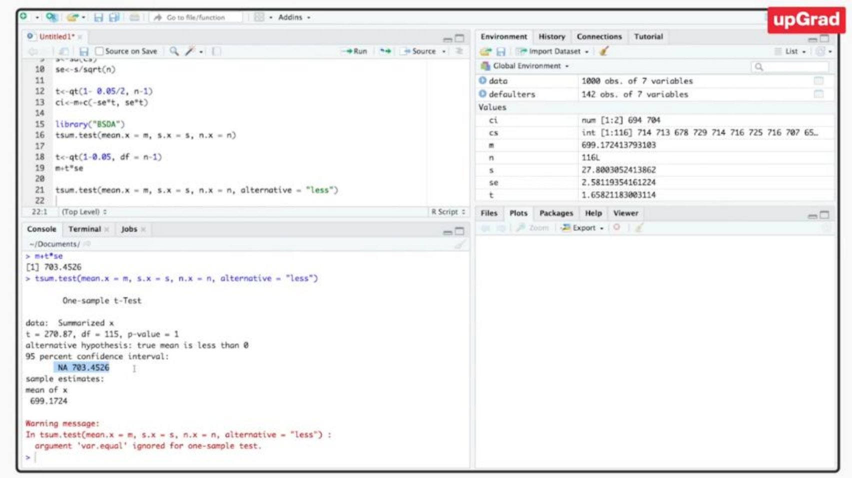


We also have the plot here, there's a 95% chance that mean credit score of defaulters lie between 694 and 704.3.



Ram and Gocha also discuss about proposing a second scrutiny or inspection for those who have a very low credit score. Further examination of such applicants can help code and called identify potential defaulters, and it can then refuse such applications.

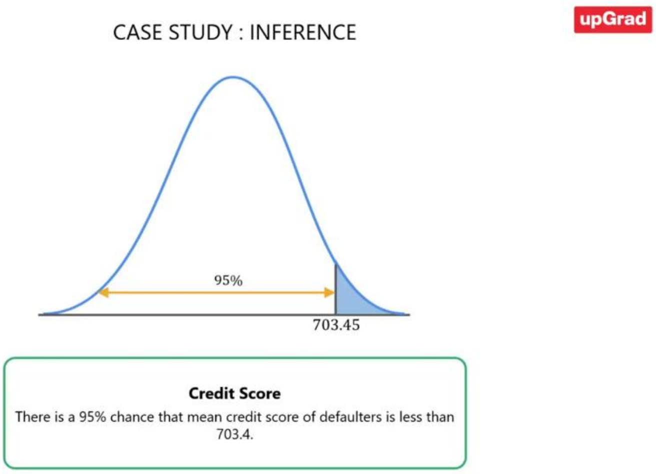
How do we define such a score or what should be the cut-off credit score for a second scrutiny? By cut-off score, we want to estimate the upper bound of credit score amongst defaulters, such that anyone below that score can be scrutinized again. In interval estimating term, we want to find one-sided confidence interval or the upper bound of credit score amongst defaulters.



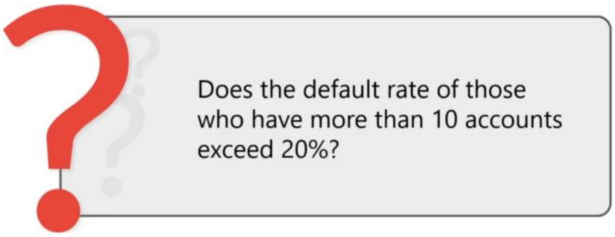
Earlier, we build a two-sided confidence interval. Now, we just want to get the critical value at the right side of the distribution. So, going back to R, first, we need the right T value at 95% confidence interval. So, we use QT function, and we pass 1 minus 0.05. Now, the 5% region is only at the right side of the distribution, and therefore, we use alpha 0.05 and not alpha by 2.

Next, we can calculate the upper bound as m+t x se. So, this is the upward bound, 703.45. With 95% confidence, population mean of credit defaulters is less than 703.4. We can also estimate confidence interval using t sum dot test function.

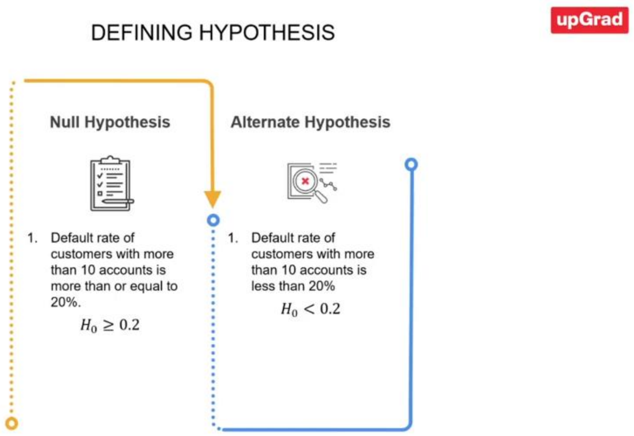
As we want to estimate the upper bound in the t sum dot test function, we will pass alternative as less. Once we run it, we have one-sided confidence interval, which is the upward bound of 703.4.



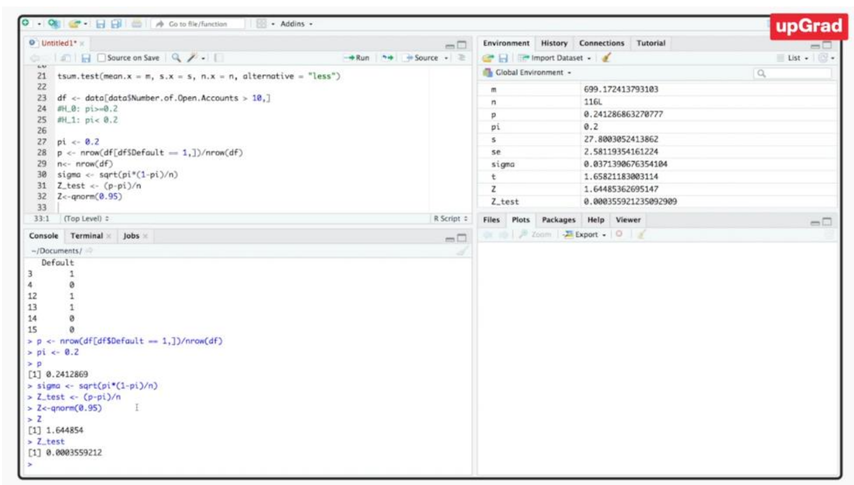
We can also see it graphically. Based on the given sample, we found that on average, there is a 95% chance that the credit score of defaulters is less than 703.4. Therefore, anyone with a score less than 703.4 should be inspected further.



Moving on, we also want to check if the population default rate of those who have more than 10 bank accounts exceed 20%. We can perform a hypothesis test for this.



So, a null would be that default rate of customers with more than 10 accounts is more than 20% or pi is greater than an equal to 0.2. And our alternative hypothesis would be default rate of customers with more than 10 accounts is less than 20% or pi is less than 0.2. This is a population proportion test. Let's do it in R.



For population proportion hypothesis test, we would want to identify the sample proportion of default amongst those who have more than 10 accounts. For that, we would subset the data and we'll define DF as data such that number of open accounts is greater than 10.

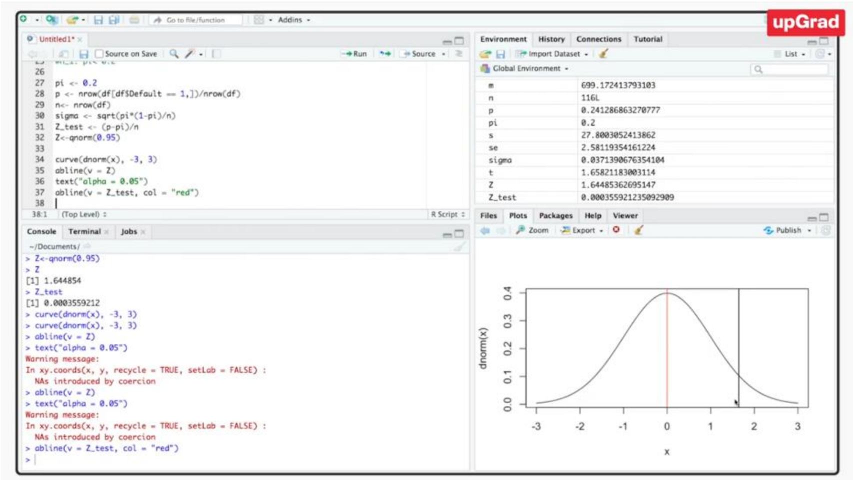
So, in DF, we have all the customers who hold more than 10 bank accounts in our sample. Based on the sample, we want to test the null or default rate of customers who have more than 10 accounts is 20% or 0.2 and the null hypothesis would be that by pi is less than 0.2. Pi is 0.2, so the pi or population proportion in the hypothesis is 0.2.

Next, we need to estimate the sample proportion or the default rate in DF. So, we can calculate sample proportion or P using this command. In the denominator, we have all the customers who are more than 10 accounts, which is the number of rows in DF. And in the numerator, we have those customers in DF who have defaulted, that is where default is equal to 1.

So, P or sample proportion of default rate for customers who hold more than 10 bank accounts is 0.24. Now to test this hypothesis, we would want to calculate sigma. We can define sample size as nrow DF, and this is how we would calculate sigma in proportion test. So, it is square root of pi x 1 minus pi divided by N.

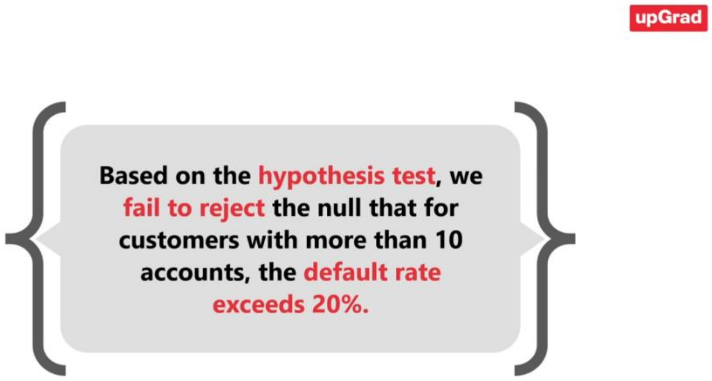
The test statistic that can be used is Z test. So, we'll define the test statistic as P minus pi divided by N, and we can calculate the critical value using qnorm function. So, we will define C as to qnorm 0.95.

This is a one-tailed test, as our alternative hypothesis is pi less than 0.2, and therefore, we use 5% as rejection region. The critical value is 1.64, while the test statistic is 0.0003, and since the Z test statistic is less than critical value, we do not reject the null that pi is greater than equal to 0.2.

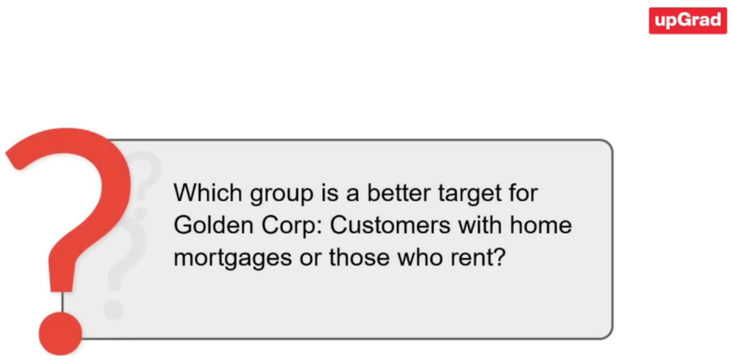


We can also plot the graph here, just remember how we will plot the graph using curve function. We will first plot the normal distribution and we will define the critical region or the rejection region.

So, this is the rejection region, and our test statistic is almost 0, it lies here. The test statistic does not line the rejection region. Hence, we don't reject the null.

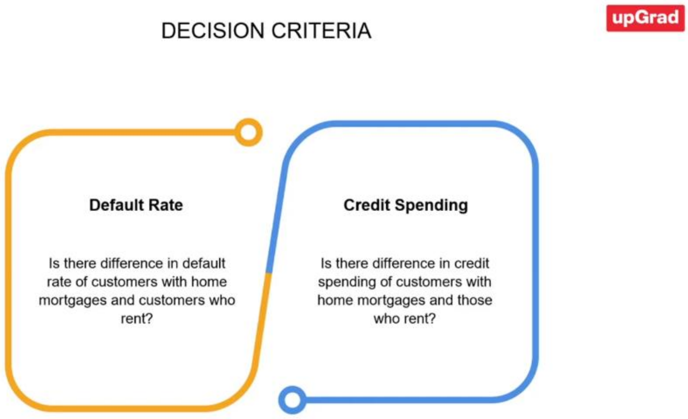


So, we can conclude that based on the given sample, we fail to reject the null that customers who have more than 10 bank accounts, the default rate exceeds 20%. Golden Corp should thus be more cautious in approving credit cards to such applicants.

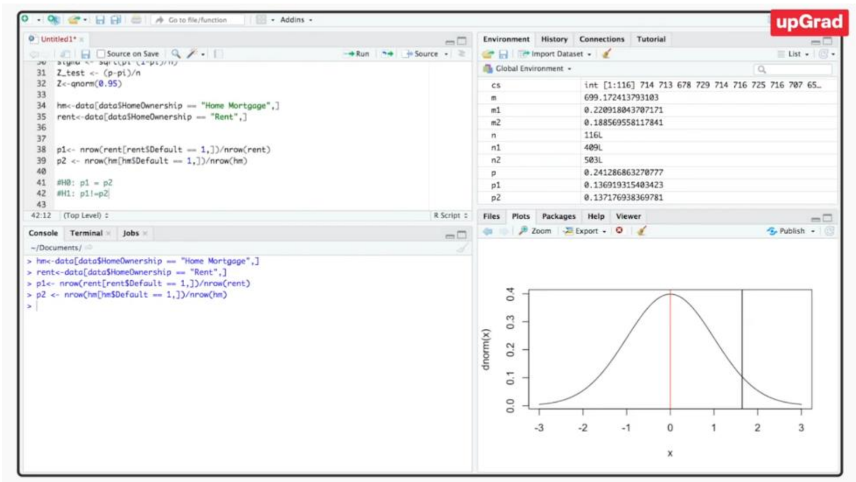


Finally, the last objective is to identify which group should Golden Corp target in its new marketing strategy. They want to identify which group can generate more revenue for Golden Corp. Is it the customers with home mortgage or is it the customers who are living on rent?

We do not include the customers who own houses here, as they are secured borrowers, and that any way given appropriate offers by Golden Corp, they want to make customized policy for the other group. The other two groups are also fundamentally comparable based on home ownership as they both pay a considerable proportion of income on houses, either in terms of EMI or in terms of rents.

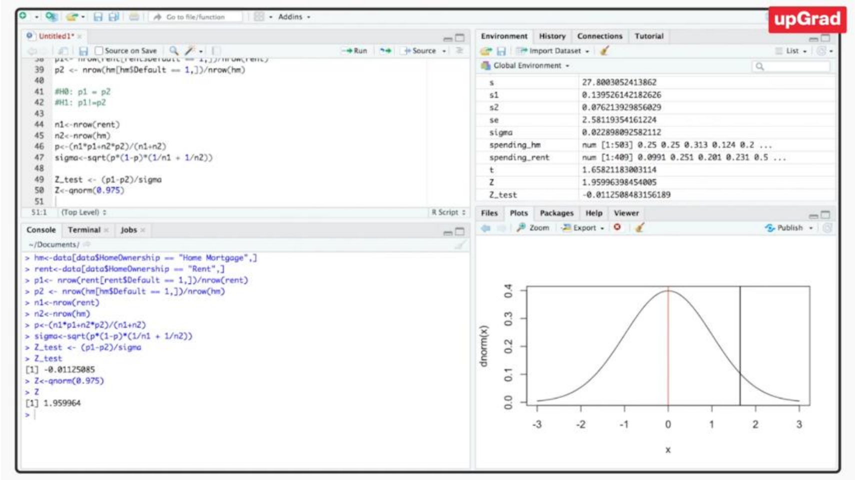


So, in order to identify the better target or the higher revenue generating group amongst the two, Ram decides to check for two factors, default rate and credit spending. So, he wants to check if there is a difference between the default rate of two groups. For this, we use hypothesis testing for difference in proportion.



So again, we will first need to subset the data to compute the proportion or default rate of the two groups. So, we'll define HM as the customers who have home mortgage, and we define rent as customers who live on rent. So, we subset on the home ownership column, and we create two subsets HM and rent here.

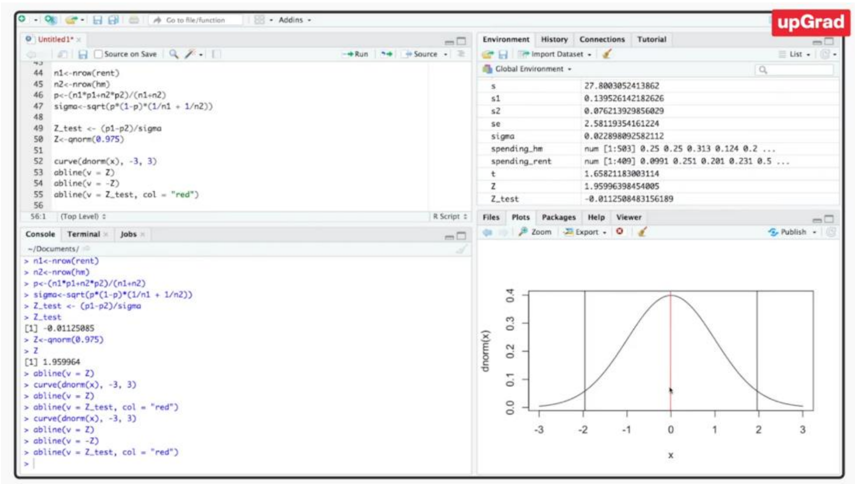
Next, we compute the default rate of each group. So, the default rate would simply be the number of defaulters in that group divided by the total number of customers and then 2. So, for example, P1 or default rate of those who rent houses would be equal to nrow, number of defaulters in the rent group divided by and nrow of rent group. And similarly, we will compute the default rate for those who have home mortgage.



Now, we want to test the null hypothesis that P1 is equal to P2, and the alternative would be P1 not equal to P2. We can define the sample size here, using again nrow command, and further for hypothesis testing of difference in proportion, we would need to compute P which is the pool proportion.

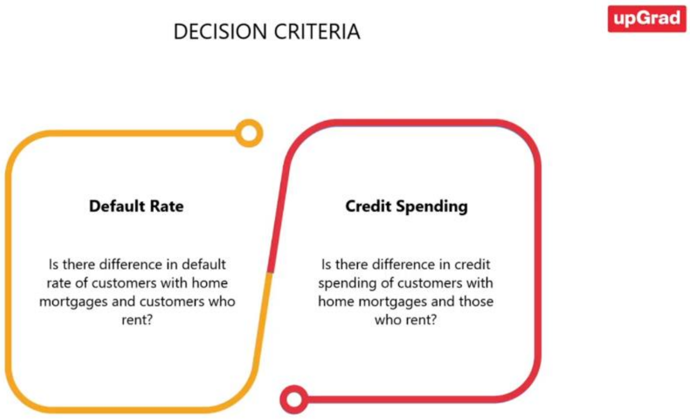
And if you remember, this is the formula for computing pool proportion, n1 into into P1 plus n2n2 P2 divided by n1 plus n2, and the sigma would be square root of P x 1 - P multiplied by 1 upon n1 plus 1 upon n2.

We have seen this formula in lectures. We have the sigma. So, we can calculate the test statistic. It would be P1 minus P2 divided by sigma. So, the test statistic is minus 0.01, and we can compute the critical value using qnorm function. Because it is a 2-tailed test, we'll use alpha by two, which is 2.5 and 1 minus 2.5 is 97.5. So, we will use qnorm of 0.975.



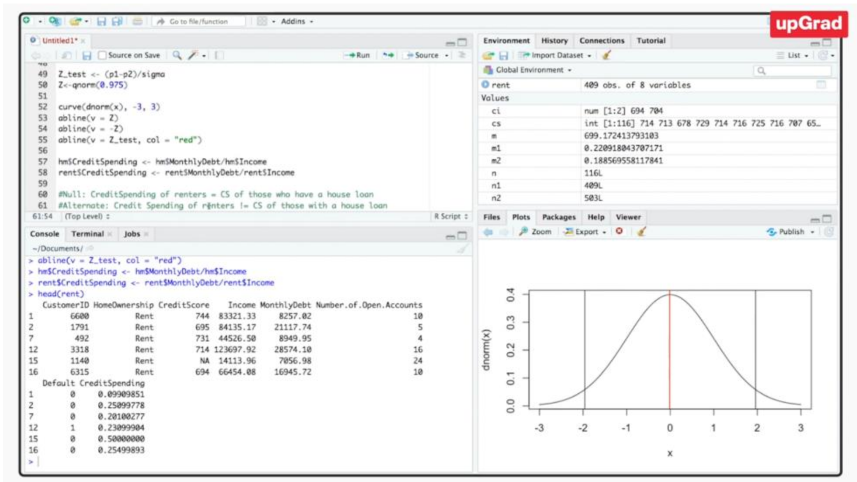
Critical value is 1.9. The test value is less than the critical value. We can also see it in the graph here. We are again using the curve function. This is a 2-tailed test. So, the rejection region would be on both side of the curves, and our test statistics lies here. It does not lie in the rejection region.

So based on the sample, we do not reject the null that default rate of customers with home mortgage is equal to default rate of customers who live on rents.



So, we have checked for the first factor default rate, we could not reject the null that the two groups have same default. Had they been different, and we could identify a group with a lower default rate, Golden Corp would have focused on that group to increase its revenue.

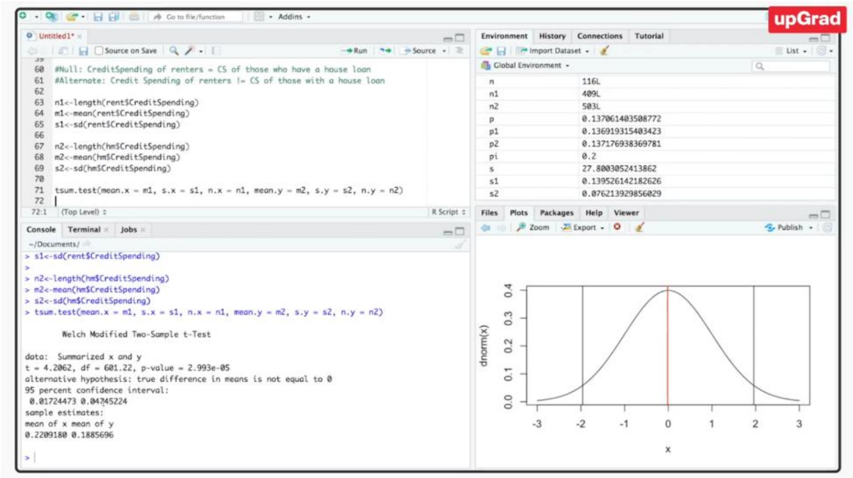
Next, we will check if there is difference in credit cards spending of customers with home loans and customers will live on rental properties. Generally speaking, customers or group who have a higher credit card bill are generally more valued by a bank or by a credit card company. It is a group which is generating more revenue for them.



So, let's now check the hypothesis if we can see a difference in credit card spending of customers who have home loans versus customers who live on rent. First, we will approximate the credit spending using monthly debt or monthly credit card bill divided by the monthly income.

So, we will define a new column in the data set called credit spending, which is the monthly credit card bill as proportion of their income. So, we define this variable in both the groups. So, we have a new column called credit spending.

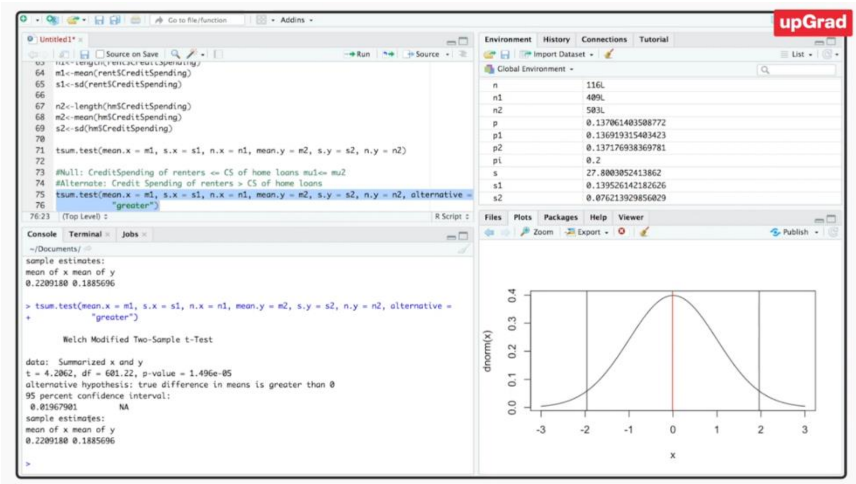
We want to conduct a difference in mean hypothesis. We want to test if credit spending of renters is equal to credit spending of those who have a house loan. So, we can define the null as credit spending of renters is equal to credit spending of those who have a house loan, and the alternative would be credit spending of renters is not equal to credit spending of those with a house loan.



Next, we can define the variables. So, we have n1 has lend of rent dollar credit spending, m1 as mean of rent dollar credit spending and standard deviation is s1, which is SD rent dollar credit spend.

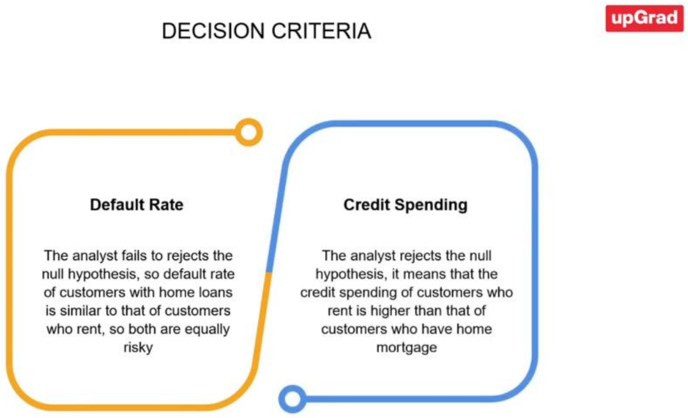
Similarly, we have n2, m2, and s2 for those who have house loan. So, we can conduct the difference in mean test using t sum dot test function, we define mean dot x s dot x and dot x as n1, s1 and n1. And similarly for the second sample, we define m2, s2, and n2.

Once you run this, you can see that the 95% confidence interval for difference in means is 0.01 to 0.04. A null is the mean of first sample is equal to mean of second sample that is mu1 minus mu2 is equal to zero. Since zero does not lie in this interval, we reject the null that credit spending of the two groups is same.

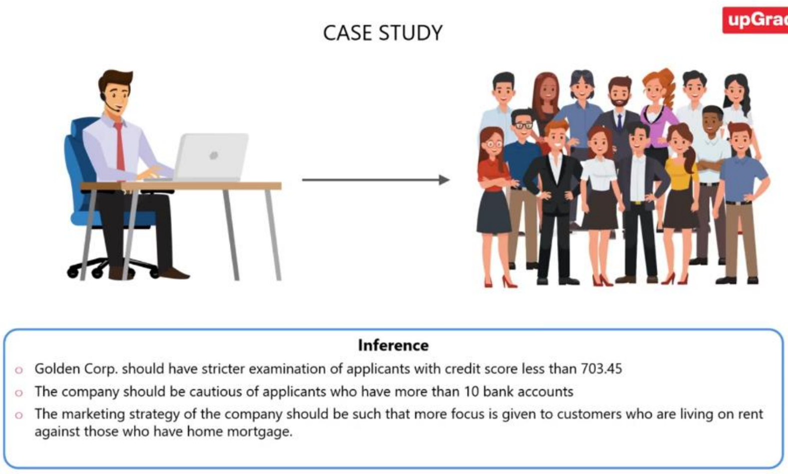


We can further test the hypothesis that spending of renters is less than spending of home loans. So, we will define the null as credit spending of renters less than equal to credit spending of home loans or mu1 is less than equal to mu2, and the alternative will be credit spending of renters is greater than credit spending of home loans.

So, as our alternative is greater, in the t sum dot test function, we'll add another argument alternative equal to greater. So, the one-sided confidence interval or lower bound is 0.01, that is, based on this sample, we are 95% confident that the difference between credit spending of renters and credit spending of home loans is greater than or equal to 0.01. Since 0 does not lie in this interval, we reject the null that credit spending of those who rent is less than or equal to credit spending of those who have home loans.



So, coming back to choosing a better target, Ram concludes that based on the above sample, he cannot reject the hypothesis that default rate between the two groups is same. Further, he is able to reject the hypothesis that means spending of those who rent is less than means spending of those who have a house loan.



Finally, based on the sample study, Ram concludes three points and reports them to Ms. Gocha.

1. Golden Corp should have stricter examination of applicants with credit score less than 703.4. We found that on average, there is a 95% chance that the credit score of defaulters is less than 703.4.

2. Secondly, Golden Corp should be cautious of applicants who have more than 10 bank accounts. Based on the sample, we fail to reject the null that default rate of customers with more than 10 accounts exceed 20%.

3. Finally, as we saw in the last test, it should target its marketing strategies to attract customers who are living on rent versus those who have a home mortgage.