Transcript

Hello everybody.

My name is Lakshmi Narayan Bharadwaj and I am here today to take you through this statistical analysis presentation done by me on health care data.

So before we jump in, please be informed that I will be switching screens from presentation slide two and our code and some other snippets.

Some other screenshots for your reference to make to help you understand the presentation better and even the data also.

Yeah, let's jump in.

So a brief introduction about what is being done in this presentation and what is the topic.

So as you can see, basically the topic here is alcohol consumption.

Alcohol consumption is not only a problem, a big problem in the UK, but in the world, so it is.

On uh.

Rising trend nowadays. So for our presentation purposes, we have considered alcohol consumption in the UK. So based on the given assignment, there were almost 338 K admissions that have happened in the hospitals of the UK.

And almost 5843 alcohol specific deaths were reported over a period of five to six years.

So in order to analyze this, we have gathered the data from health survey for England.

In the year.

2011 for our data analysis.

So here in the presentation if you can see so this is the object that I have inserted here in each and every slide.

So for our reference.

So basically what happens is if you try to double click this.

There is no sign.

Schedule classes happen.

We will be able to basically read the data or look at the data from here here itself.

There is no need for us to again go to the location where the file is located and then open it from there.

So we can basically open it from here.

Here in this case this is the.

2011 data which we have, which we have gathered, which was given to us.

And this is basically in dot Avi format so.

In order to open this file, there are are.

No software readily available using which this data can be opened.

However, you can view this data in our studio after importing it.

I will basically try to show you.

What do I mean by that?

So here if you can see I have read the data from my local machine into into.

This ask to be environment and if you wish to look at it all you have to do is click here in the environment variable section.

So basically you can see all the data that is basically you know.

Present in that visit in that.

So this is how savvy formats are red.

And now jumping back to the presentation, so this is what the data we have considered for our analysis purposes.

So going on to the next slide.

We were asked to perform some descriptive statistics on the data.

So basically there were a few questions which which which I have included here in this presentation slides.

So the first one.

Uh is saying how many people are included in the sample. So here for our data analysis purpose, I'd like you to inform one more thing that we have considered a sample of the total population of the data. So basically our data had about 10,600 rows out of which I have.

Considered 7000 rows.

For our data analysis purpose, there are 58 columns in total and I have included all the 58 columns for for our analysis. So the first question in the descriptive statistics section says how many people are included in the sample.

In order for us to.

Determine that we have used a sample command in R so.

Basically, if we try to try to use the sample command in our, this is what the output looks like.

So see, right now I'm going to double click this object here for you.

It's gonna ask you.

PowerPoint is going to ask you that.

Do you want to open this file?

So since this slide is safe, you can open it.

So here you go. If you look at this sample data we have stored, might we have stored 7000 rows of our whole data into a sample data variable?

So that is basically sample data consists of 7000.

Rows in our data.

So that is.

How many rows that that's how many rows are included in our sample.

For our data analysis purpose.

So moving on to the next slide.

I I'd also like you to apologize for not presenting you in the presentation format because I would not be able to open up the objects and show it to you while while presenting and explaining it while presenting in explaining.

The presentation, yeah.

So the second bit in the descriptor stats was what is the percentage of people who drink alcohol.

So basically there are there are there variables in this data set like gender.

How much do they drink?

Well, how much?

How many units of alcohol do they consume?

What is their household income?

And so on and so forth.

So basically here what what is required for us is basically the gender which is male and female.

And how much percentage do they consume?

How much percentage does each individual consume?

So in order for us to determine this, we have done something called.

We have used some library called as G models.

There is a library in our called G models that contains across.

Table function a function called as cross table.

So basically if you execute this cross table function by giving in whatever the input we want, we will give the.

We will get the output in such manner.

Once again, I'll show you what happens.

So when you run.

When you run the cross table command so this is what happens.

So sample data.

We are performing our analysis on sample data.

Here sex is basically the gender one and two which is 11 is considered as female and two is considered have made sample data drinking as of now.

So sample this column basically constitutes whether a person drinks or not.

So if we look at this, we get an output.

In the form of a matrix, right?

If you look at it, yeah this is a matrix form of output.

So basically.

Based on our question, what was that question?

What is the percentage of people who drink alcohol?

So let's say if one is the number of people who drink alcohol and two is the number of people who don't.

So the base basically total number of people here is 4405. This 4405 is obtained by adding 2071 and 237.

2334.

Which which one converted into percentages is about 78.6%.

So out of our given population, out of a given population of 7000 rose nearly 78.6% of.

People consume alcohol.

So this is the second bit of the descriptive stats.

Moving on to.

The next slide now.

What percentage of women?

In the sample, what percentage of women are there in the sample?

So in order to find out that we have again used the cross table function calculate the result.

So if we look at the matrix here, I'll directly show the output.

If you look at the way if you look at the matrix, here women in sample we have used the sample data dollars sets which is the gender basically so out of which.

Let's assume one is female and two is made for the set variable. In that case we have 3231 females, which is about 46.2% of the total 7000 population.

So that is the number of females that are present in our considered data population.

So yeah, that is a.

That is the next bit in the descriptive stats.

Moving on to the next next question, in the descriptive statistics bit, what is the highest educational level?

This is a pretty straightforward question.

So what we have done here is we have stored the output from sample data into a variable.

This variable highest level.

And we have simply executed the the variable to get our results.

So if we look at it, if we look at the result, this is how we will get the result.

Highest education level is given by sample data.

Dollar top, quad three top quad three is basically the variable that gives you the highest education level in this format.

If you execute this you get.

The output in such manner.

So the value we have values ranging from minus nine 2 + 7 and you have labeled for each label has its own you know description.

So for minus nine it is refused minus 8.

This don't, no.

So if you look at it for for us to answer our question we require one which is North VQ4 and Q5 degree or equivalent.

So based on the above command one is the highest educational level which is VQ4 and Q5 degree or equal which is the highest.

Form of degree available in this.

A label in this label section in this link.

In this label set if you can call it yeah.

So that is the next bit.

Moving on to the next question, in our descriptive stats fit.

What is the percentage of divorced and separated people?

So there are two calls.

So there are.

So there's a column in our data set known as Mars Stat.

See if you can see it here.

That gives you basically again, levels of data levels in which.

The data is spread spread across, so then in order to obtain the.

Percentage of divorced and separated people.

What we have done is basically determine.

The levels for.

In the data, then use crossed able to find out the percentage.

So let's see we have two outputs here.

Let's see.

So first we will see the levels.

So if you if you execute this command in our it will give you the levels.

So they are again ranging from minus 9 to 7 in which four gives you the separated 5 is divorced which we want for our for our analysis purpose.

So here 4 is separated, 5 is divorced.

Again, if you look at the output here percentage.

You will get the same. So for four we have 0.138 people who are separated and for five divorced are these many people which if converted into percentage is is this much?

2.4% of people and 6.9% of people are separated and divorced respectively. So this is how this is how we are able to interpret this bit.

So moving on to the next and final bit of the descriptive stats.

So we were asked to calculate all these summary statistics.

Mean, median, mode, minimum, maximum range, and standard deviation of.

These three variables household size, BMI and Azure last birthday.

So we have tabulated the results as follows.

One keen thing to observe here is mode when calculating when calculated for the BMI variable.

Give the following output.

So basically there are basically.

These many number of instances which have repeated the same number of times.

So that is the reason.

Why I have?

Screenshot I've put the screenshot for mode mode value for the BMI variable.

And yeah, these are the commands that are used to obtain the output, so yeah.

The code is pretty self-explanatory if you look at it. If you look at it.

So that is how we have obtained the summary statistics of all three variables.

So this is about the differential descriptive statistics.

We now move.

On to the inferential statistics bit of our assignment.

Basically inferential statistics.

Is something that are that is used to establish you know a deeper relationship with in the variables of any data.

So basically what we do here is we design a null hypothesis, something called as null hypothesis, which we basically try to reject based on.

Whichever test we are, we are running and whichever.

Whatever P value, whatever the static value we obtain after running the test.

So in the inferential statistics bit, we were asked to run a significance test to find out which gender drinks more alcohol.

So if you observe.

Male and female are two categorical variables and in order for us to determine.

Whether a gender, whether gender plays a role or not, we have done the following so.

Once you open this.

We have basically cross tabulated the gender and people who drink, basically who drink or not and then we have converted it into a table or tablets or tabular form.

After converting it into the tabular form, we have run the chi squared test because it works well on categorical data.

Chi Square test works well on categorical data.

So then if we interpret the test, if we look at it, we get the X squared value.

Is 89.97 degrees of freedom is 1. The P value is way way less than 0.0.

So then in this case we can basically.

Evidently reject the null hypothesis.

Yeah, so by saying that gender does not play a role in consuming alcohol.

So that is how we interpreted this bit, moving onto the next bit.

Runner significance test to find out which region drinks more alcohol.

This the process is similar only only change here is instead of sex we we've introduced the region.

So by interpreting the code what I've observed was is religion had.

Nine different levels.

If you observe, yeah, you can see levels from one through 9 and if a person drinks now or drinks or not.

So that is one or two.

If we assume one through 9 as regions for GR1, this is a variable that is basically you know used to determine the region.

And one and two for drinks and.

Not drink alcohol.

We can see that eight region 8 have more number of people, almost 744 or 745 people who consume alcohol.

So that is how you can basically determine.

And one more again we have determined in order to interpret this result, we have performed a CHI square test.

Which gave us a value X squared value of 65.98 degrees of freedom because yeah 129 there are, yeah 8 degrees of freedom and P value is way lesser than I get zero point.

05 meaning we can reject the null hypothesis saying region also does not pay a role, does not play a role in consuming alcohol.

So this is the 2nd between the inferential statistics part.

Moving on to the third.

But so we were we were asked to investigate whether there is a statistical difference between men and women on the following variables.

In order to do so, we must basic firstly form a null hypothesis for both these variables, palette heightened valid wait and then run an independent T test on both of them and then obtain the value obtained, the P value and interpret the result so.

Which is what we have done here.

So basically the null hypothesis of both the variables.

There is no statistical difference between men and women on valid height.

And there is no statistical difference between men and women on valid wait are the two null hypothesis that we have designed?

So in order for us to reject, we have done the the independent T test.

The outputs of which are shown here.

So see yeah below other T tests for both height and weight. Ballot height is this one. So if you look at it, the T statistic is 21.455, P value is 1.

We have obtained P value as one here. So if you look at the confidence interval also it ranges from minus Infinity to 11.0.

So based on the obtained P value, we fail to reject or we accept the null hypothesis.

So in this case it stands correct that there is no significant difference between men and women on valid height.

Similar is the case for valid wait also.

So we have done the test on valid weight also.

And then interpreted the result obtained the P value as one T statistic as 14 and basically again null hypothesis stands correct here.

So we fail to reject the null hypothesis by saying that there is no significant difference between men and women on valid weight also.

That is one.

That is the next bit that we have done.

Moving on to the correlation part, what is correlation?

Basically, correlations is nothing but a measure that gives you.

How well?

Or badly.

2 variables in a given data set are related.

I hope you follow.

So if you look at it here, we have basically tabulated our correlation.

We were asked to tabulate the correlation between these these four variables.

Person who drinks nowadays?

The total household income.

Age at the last birthday and gender.

So we have basically tabulated the results in this manner.

If you wish to look at the code, this is how we have obtained the correlation matrix.

If you call it, yes.

So there is directly a function in our that can give out.

The correlation between any variables in given data set.

So this is how.

We have tabulated.

Uh, also, there's there's a few more points to know here that correlation basically ranges from -, 1 to positive one whenever a correlation value is closer to -, 1.

We say that.

There is a negative correlation between any given.

Two variables.

Let me explain using this table here.

For instance, let's let's compare the total household income.

With the age of a person. So if you look at it, the value is about 0.047 which is kind of which is a positive variable which is a positive number for sure.

But it is not close to plus one.

So in this case we say that the total household income.

Is positively correlated with the age of a person.

And in a similar way, let's now consider a negative value.

If if we consider the gender and the birthday, if you look at it, the value is -, 0.008 which is there is a negative correlation.

Very slight negative correlation between gender and the age.

Of a person.

So that is how and if you look at.

If you look at, if you try to compare person drinks nowadays and person brings nowadays, the value will result to one because you're basically comparing 2 same variables, so it is bound to give you one.

That is how we have interpreted the correlation table here for the given variables.

I think we are right now at the end of our statistics, inferential statistics bit we now move on to.

To some literature that some literature research that I have done based on gender and alcohol consumption in the UK.

So these two papers, 1 written by Mr Moinuddin al in 2016, they have shown that alcoholism was on the rise due to factors like gender, SES, like the economics.

Status, culture shift etc in their article alcohol consumption.

So this is the title of their article alcohol consumption in gender Critical Review you can access.

I have included a hyper link here.

You can access the article here.

Yes, in a.

Similar way, Sambhal Sambhal al in 2020 have argued that alcohol consumption is more hazardous than illicit drug consumption among young gutters. So.

This article as team has focused on a specific group of others, specific group of age group of adults and have conducted statistical analysis techniques to determine whether alcoholism is present or not in the said age group in the set.

Yeah, age group.

So these are the two literatures that have.

That are paved way that have you know.

That were very very useful for me to compare my analysis what what has happened in theirs?

So in conclusion, both the articles have highlighted that.

Males are more likely to consume alcohol than females.

However, Moinuddin paper considered factors like region, education, cultures, etc.

To understand the trend in female alcohol consumption.

Similarly, son Val have also considered factors like ethnicity, smoking habits, physical.

To determine the rice.

In alcohol consumption in young adults, irrespective of the gender.

So this is how we have concluded the paper by by going through them by observing both articles.

I would like to recommend that if stringent policies and regulations were enforced on alcohol induced.

You know, there is a chance to curb the rising consumption of alcohol in the world, so it will also be more beneficial if we create awareness amongst the population regarding the hazards that alcohol can cause.

You know, that is one more recommendation that I'd like to provide from my side.

So this I hope.

But I have made sense with my presentation and these are the references that have been included the tables that I've used and the append is present in the presentation.

So thank you for your time.

I really appreciate it.

Yeah. Thank you.