Name: Pratik Premkishor Mantri

NUID:001353214

Module2

Appendix:

|  |  |
| --- | --- |
| Index | Page no |
| Introduction | 2 |
| Data analysis | 2-8 |
| Key-points | 9 |
| References | 9 |

**Introduction:**

In this dataset of baseline survey, we have 13 columns and 256 rows. This data is about undergraduate students, their age, their majors, their performances like how proficient they are in R programming language, Probability and distribution, statistical analysis, how much experience they have and many more.

**Data analysis:**

We do data analysis to get some crucial insights in dataset. We have plotted some graphs and statistical tables which can give proper insights. To get insights inside the data I have plotted different graphs such as jitter plot, scatterplot, histogram, bar plot etc.

Firstly, we will see how many students are enrolled in different undergraduate major:

Chart, bar chart

Description automatically generated

From the above bar plot, we can say that students majoring in computer/IT field are highest and less in chemistry/biology major.

Jitterplot:

The usage of jitter is a best technique in box plot or scatter plot. Jitter is nothing but a random value (or for our purposes pseudo-random) which is assigned to the dots to separate them so that they should not be plotted directly on top of each other.

Chart, scatter chart

Description automatically generated

In the above jitter plot, we can take one categorical and two numerical.

Here we can see that, only peoples from Major: Computer/IT, Engineering, and other have an experience more than two years in r language. we can conclude that these Major are related to the R language. It looks like these major courses are dependent on R language. We can also see here that most students are having work experience of max two years. Almost more than 60% of work experience comes from majors: computer/IT, Engineering, finance, and others.

Let’s do more analysis and see in which field we use R language most:

Chart, scatter chart

Description automatically generated

The majors such as Computer/IT, Engineering, Finance, and other are more leaning towards R language. Most of the students looks Beginner in R only few students have an expertise in the R Language.

Lets now see whether AGE matters in the work experience and R programming expertise:

Chart, scatter chart

Description automatically generated

The use of three-line table format:

Table

Description automatically generated

The normal tables are used **t**o organize detailed or complicated data, which is described frequently in the dataset, which helps user reader to quickly see the results and patterns in the data. They are **used** to highlight trends or patterns in the dataset.

Here in the above three-line table format, we can read and explain the complicated data easily. We even can make some interpretations like stat and analytics experience is highest in major computer/It, which is two years.

Here, we can see that Age is directly proportional to the work experience of the student which is obvious but the interesting observation is there is a decline in expertise as the age of students increases. Trend line is going downwards.

Chart, bar chart

Description automatically generated

There are very less people who literally have expertise in R language. Here the interesting most of the student knows R, they might be beginner but have some experience with R except. We can say from observation that R is used in almost every major.

Chart, scatter chart

Description automatically generated

Here, we can see that R language plays a vital role in stat and analytics. The trend line shows us that more the year of studying statistics and analytics, more you learn R-language. Hence become expert in R language.

Lets explore other fields:

Chart, histogram

Description automatically generated

From the above observation we can say that most students are experts in central tendency concepts, and less expert in probability and bivariate concepts. it may be the possibility they are more interested in this particular concept or they have been taught nicely etc.

Lets get some insights by plotting a graph of Undergrade major :

Chart, bar chart

Description automatically generated

From the above observation we can say that the most students from the geology/GIS major who are experts at Central tendency concepts, and the major ‘Other’ have zero experience in central tendency concept. We can make prediction that central tendency and variance plays important role in major ‘Geology/GIS’.

Box-plot:

Box-plot gives us five summary about the data.

1. minimum: Which is lowest normal point in the data.

2.Q1(25th percentile): It is nothing but the sub data which belong to the 25th percentile of whole data.

3. Median: Is the median of the data, it can be calculated as. For example: For even numbers. 1,2,3,4 then the median is 3+4/2=3.5 and if the value of data point is odd then 1,2,3,4,5 then 3 would be our median.

4. Q3(75th percentile): It is nothing but the sub data which belong to the 75th percentile of whole data.

5. Intermediate range (IQR): The range between Q1 and Q3.

6. Maximum: The highest normal point in the data.

The outliers are values which are out of the range of Normal values. Sometimes they have been misplaced by data-entry people.

Chart, box and whisker chart

Description automatically generated

Here we can see, the maximum age of the student belongs to the Major “others” and have a median age of 25. In the major ‘chemistry/biology and telecommunication the range of age of students are 25 and 22 respectively.

**Key points:**

* we can say that students majoring in computer/IT field are highest and less in chemistry/biology major.
* only peoples from Major: Computer/IT, Engineering, and other have an experience more than two years in r language.
* Almost more than 60% of work experience comes from majors: computer/IT, Engineering, finance and others.
* The majors such as Computer/IT, Engineering, Finance, and other are more leaning towards R language.
* Age is directly proportional to the work experience of the student, which is obvious, but the interesting observation is there is a decline in expertise as the age of students increases.
* R programming language is used in almost every major.
* There are very less (approx. 10%) students who literally have expertise in R language.
* that R language plays a vital role in stat and analytics. The trend line shows us that more the year of studying statistics and analytics, more you learn R prog. language. Hence become expert in R language and vice-versa
* most students are experts in central tendency concepts, and less expert in probability and bivariate concepts.
* most students from the geology/GIS major who are experts at Central tendency concepts, and the major ‘Other’ have zero experience in central tendency concept. We can make prediction that central tendency and variance plays important role in major ‘Geology/GIS’.
* Age of engineering major student the smallest which is 21 and age of major “other” is highest which is 27.
* Most students have experience in R language than the other work experience so we can see student have started interest in R language so universities should start new classes for R-language.

**References:**

K, M. (Director). (2019, October 15). *R-studio tutorial: Creating fancy jitter plots with ggplot* [Video file]. Retrieved March 20, 2021, from <https://www.youtube.com/watch?v=g6d1V4x_U7E>

Slutsky, D. (2014, November 12). The use of tables. Retrieved March 20, 2021, from <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4208953/>