Fyez Dean
Darren Yu

TIM 147 Homework 1

Summary of data:

From our usage of box charts, we found that there is nearly equal amount of college and non college educated customers. Their incomes range mainly from the 40 to 60 thousand dollar mark. As for leftover minutes, the greatest frequency is marked around near zero for the box charts; however the most amount of leftover minutes pertains before the 20 minutes mark and significantly drops off. As for customer housing, customers were likely .1 and .3. Regarding the cost of their phones, most devices ranged from 100 to 400 dollars and the price held relatively constant towards the higher end with a frequency of 500 towards the end of the phone price brackets in the data set. Analyzing call lengths, most calls did not exceed 15 minutes; and were typically only five minutes over and calls beyond that period had a low frequency below a 100 from ten minutes and onwards. Average calls typically were sub five minutes with a small minority of 8 to 15 minutes. Most customers felt unsatisfied or very unsatisfied and people with positive experiences were extremely low. In fact, most customers are either switching plans or considering switching. However, usage suggests that customers typically use very little or little or only a lot. There doesn't seem to be any middle ground in regards to usage.

Significance of data:

The analysis of our customer data reveals some key insights. We have a nearly equal distribution of college and non-college educated customers, primarily earning between $40,000 and $60,000 annually. When it comes to leftover minutes, the majority of customers use very few, with the most common range being under 20 minutes.

In terms of housing, most of our customers fall within the 0.1 to 0.3 range. Phone costs are concentrated between $100 and $400, and we see a slight uptick in the $500 range.

Analyzing call lengths, most calls are brief, with the majority lasting less than 15 minutes. Very few calls exceed this duration, and the average call duration is around five minutes.

Customer satisfaction is a concern, with most customers reporting dissatisfaction. Positive experiences are rare. The data indicates that many customers are either switching plans or considering doing so.

Usage patterns show that customers tend to use very little or a lot, with a lack of a middle ground.

In summary, the data suggests that we have a diverse customer base with income, housing, and phone cost preferences, but there are challenges in terms of customer satisfaction and retention. It's essential to focus on addressing these issues while recognizing the usage patterns of our customers.

Histogram:

Here are the listed reasons as to why the Histogram is useful when representing Data.

1. **Data Distribution Visualization:** Histograms provide a visual representation of the distribution of data. They allow you to see how data is spread across different ranges or bins, making it easier to understand the central tendencies, variations, and patterns in the data.

2. **Identifying Patterns and Trends:** Histograms help in identifying patterns or trends in data. For example, you can see if the data is normally distributed, skewed to the left or right, or has multiple modes. This information can be valuable in making data-driven decisions.

3. **Outlier Detection:** By examining a histogram, you can quickly spot outliers—data points that fall significantly outside the main distribution. This is crucial in identifying data anomalies that might require further investigation.

4. **Data Summarization:** Histograms provide a concise way to summarize a large dataset. Instead of examining individual data points, you can get a sense of the data's characteristics by looking at the shape and characteristics of the histogram.

5. **Choosing Appropriate Statistical Analysis:** Depending on the shape of the histogram, you can choose appropriate statistical tests and methods for analysis. For example, a normal distribution might suggest the use of parametric tests, while a skewed distribution might lead to non-parametric tests.

6. **Data Preprocessing:** In data preprocessing, histograms can be used to determine suitable bin sizes or intervals for discretizing continuous data. This is particularly important when dealing with data for tasks like classification or clustering.

7. **Comparison Between Datasets:** You can easily compare the distributions of different datasets by overlaying their histograms. This can be useful when you want to assess whether two or more datasets have similar characteristics.

8. **Data Quality Assessment:** Histograms can reveal data quality issues, such as missing data or data entry errors. If a histogram looks strange or inconsistent, it may be an indication that further data cleaning is necessary.

9. **Communication:** Histograms are a powerful tool for communicating data insights to a non-technical audience. They offer a clear and intuitive way to convey information about data patterns and distributions.

10. **Decision-Making:** When making decisions based on data, understanding the distribution of the data through histograms can help in risk assessment, resource allocation, and strategy formulation.

Box Plots:

Box plots, also known as box-and-whisker plots, are a valuable graphical tool in statistics and data analysis for several reasons. They provide a concise and informative summary of the distribution and key statistics of a dataset. Here's why box plots are useful:

1. **Summary of Data Distribution:** Box plots provide a quick and clear summary of the distribution of a dataset, allowing you to see the central tendency, spread, and overall shape of the data.

2. **Outlier Detection:** Box plots help in identifying outliers, which are data points that fall significantly outside the main distribution. Outliers can be easily spotted as individual data points beyond the "whiskers" of the plot.

3. **Central Tendency:** The median (or second quartile) is represented by the line inside the box. It provides a robust measure of the center of the data, making it resistant to extreme values or outliers.

4. **Spread of Data:** The interquartile range (IQR), which is the range between the first quartile (Q1) and the third quartile (Q3), is represented by the width of the box. It gives a sense of the spread or variability of the middle 50% of the data.

5. **Skewness and Symmetry:** By examining the position and length of the whiskers relative to the box and the median, you can infer whether the data is symmetric or skewed, and in which direction.

6. **Comparison of Distributions:** Box plots are useful for comparing the distributions of multiple datasets side by side. This is particularly helpful when you want to assess how different groups or categories of data differ in terms of central tendency and spread.

7. **Non-parametric Data:** Box plots are especially useful for visualizing the distribution of non-normally distributed data and data that may not meet the assumptions of parametric statistical tests.

8. **Data Preprocessing:** When dealing with skewed data or outliers, box plots can help you decide on appropriate data preprocessing steps, such as transformation or outlier removal.

9. **Communication:** Box plots are easy to understand and are useful for communicating data characteristics to a non-technical audience, such as stakeholders or decision-makers.

10. **Data Quality Assessment:** Unusual features in box plots, like extreme outliers or non-symmetrical shapes, can indicate data quality issues or data anomalies that require further investigation.

11. **Decision-Making:** When making decisions based on data, a quick look at box plots can provide insights into the variability and central values of the data, aiding in risk assessment, strategy formulation, or resource allocation.

Sources

Chat-gbt prompts

From our usage of box charts, we found that there is nearly equal amount of college and non college educated customers. Their incomes range mainly from the 40 to 60 thousand dollar mark. As for leftover minutes, the greatest frequency is marked around near zero for the box charts; however the most amount of leftover minutes pertains before the 20 minutes mark and significantly drops off. As for customer housing, customers were likely .1 and .3. Regarding the cost of their phones, most devices ranged from 100 to 400 dollars and the price held relatively constant towards the higher end with a frequency of 500 towards the end of the phone price brackets in the data set. Analyzing call lengths, most calls did not exceed 15 minutes; and were typically only five minutes over and calls beyond that period had a low frequency below a 100 from ten minutes and onwards. Average calls typically were sub five minutes with a small minority of 8 to 15 minutes. Most customers felt unsatisfied or very unsatisfied and people with positive experiences were extremely low. In fact, most customers are either switching plans or considering switching. However, usage suggests that customers typically use very little or little or only a lot. There doesn't seem to be any middle ground in regards to usage. condense and write summary in the perspective of a manager reading a business repoty

Response:
Summary:

The analysis of our customer data reveals some key insights. We have a nearly equal distribution of college and non-college educated customers, primarily earning between $40,000 and $60,000 annually. When it comes to leftover minutes, the majority of customers use very few, with the most common range being under 20 minutes.

In terms of housing, most of our customers fall within the 0.1 to 0.3 range. Phone costs are concentrated between $100 and $400, and we see a slight uptick in the $500 range.

Analyzing call lengths, most calls are brief, with the majority lasting less than 15 minutes. Very few calls exceed this duration, and the average call duration is around five minutes.

Customer satisfaction is a concern, with most customers reporting dissatisfaction. Positive experiences are rare. The data indicates that many customers are either switching plans or considering doing so.

Usage patterns show that customers tend to use very little or a lot, with a lack of a middle ground.

In summary, the data suggests that we have a diverse customer base with income, housing, and phone cost preferences, but there are challenges in terms of customer satisfaction and retention. It's essential to focus on addressing these issues while recognizing the usage patterns of our customers.

Chat gbt prompt:

how to import xlsx to google colab

To import an XLSX file into Google Colab, you can use the following steps:

Upload the XLSX File to Google Colab:
You can upload the XLSX file directly to Google Colab from your local machine. Here's how you can do it:

- In your Google Colab notebook, click on the "Files" tab in the left-hand sidebar.
- Click the "Upload" button and select the XLSX file from your computer.

Mount Google Drive (Optional):

If you have the XLSX file stored in your Google Drive, you can mount your Google Drive to access the file. This is an optional step and can be useful if you want to work with files stored in your Google Drive. Here's how to mount your Google Drive:

python

Copy code

```
from            import
           '/content/drive'
```

Follow the prompts to get the authorization code and authenticate your Google Drive.
Read the XLSX File:

Once you've uploaded the XLSX file, you can use a library like pandas to read the data from the XLSX file. If you haven't already installed pandas, you can do so with the following command:

python

Copy code

Then, you can read the XLSX file into a Pandas DataFrame like this:

python

Copy code

```
import       as

     'your_file.xlsx'
```

Replace 'your_file.xlsx' with the actual file path or name.
Verify the Data:

You can use the head() method to display the first few rows of the DataFrame to verify that you have successfully imported the data:

python

Copy code

That's it! You've now successfully imported an XLSX file into your Google Colab notebook and can start working with the data using Python libraries like Pandas.