DOCUMENTATION

Dataset:

This dataset has details of 1000 users from different backgrounds and whether or not they buy a bike. This data can be used to build the dashboard in Google Sheets. There are some NA (Null / Empty) values injected in the dataset. Use this dataset for Data Cleaning, Exploration, and Visualization.

Columns:

The Dataset contains the following columns. The target variable is 'Purchased Bike'.

- ID
- Marital Status
- Gender
- Income
- Children
- Education
- Occupation
- Home Owner
- Cars
- Commute Distance
- Region
- Age
- Purchased Bike

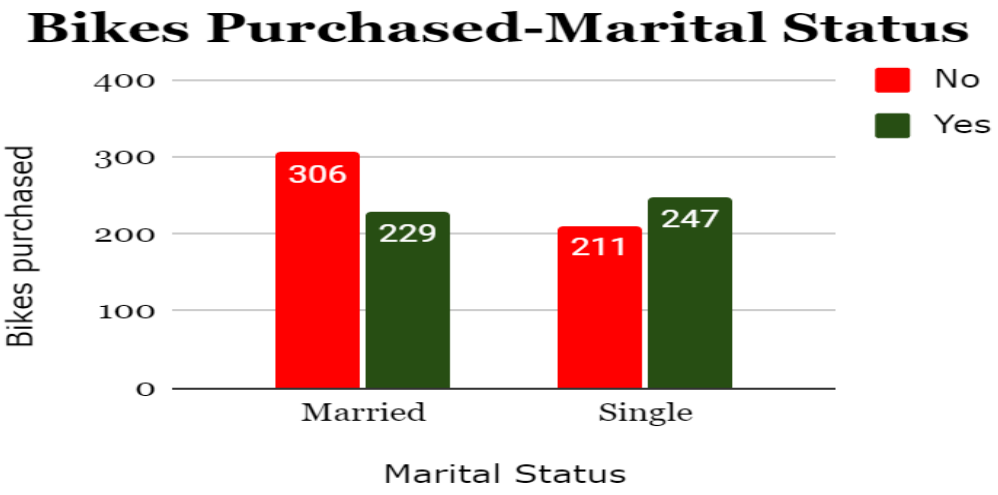
Bar Chart (Marital Status):

Question: How does the count of bike purchases vary among different marital statuses? Are married individuals more likely to purchase bikes?

Step-1: Create a pivot table having columns with "**Marital Status**" and rows with "**Purchased Bike**", values with "**Marital Status**", Filter Null values in "**Marital Status**".

Step-2: Select pivot table, select insert option and insert a chart. By default, "**Bar Chart**" is inserted.

Step-3: Customize the plot by changing **colours, legend, serie**s etc to understand the plot easily by the user.



- Conclusion: Single people are purchasing Bikes more than Married people.

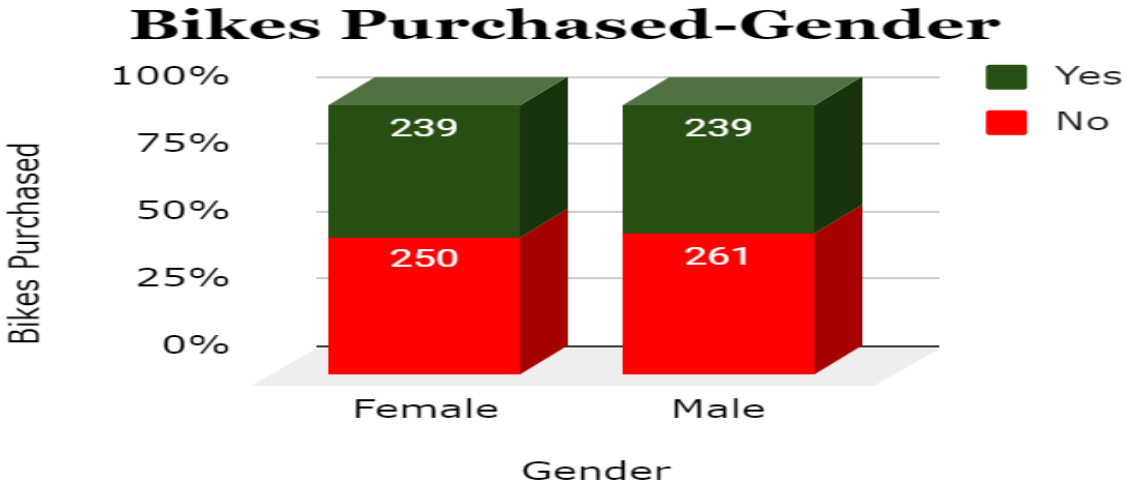
Bar Chart (Gender):

Question: Build a bar graph to compare the count of male and female customers. Does gender influence bike purchases, and if so, to what extent?

Step-1: Create a pivot table having columns with "**Gender**" and rows with "**Purchased Bike**", values with "**Gender**", Filter Null values in "**Gender**".

Step-2: Select pivot table, select insert option and insert a chart. By default, "**Bar plot**" is inserted.

Step-3: Customize the plot by changing **colours, legend, serie**s etc to understand the plot easily by the user.



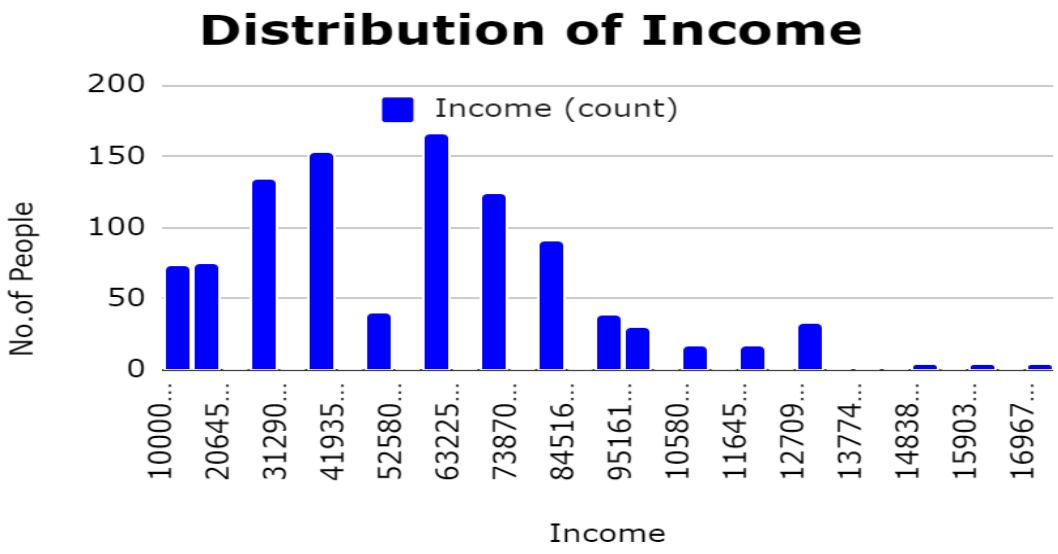
❖ Conclusion: There is equal competition in Purchasing bikes between male and female. But Male are not purchasing more than Women.

Histogram (Income):

Question: What is the distribution of income among bike buyers? Are there specific income brackets that show a higher likelihood of bike purchases?

Step-1: Select the column "**Income**" and insert a Chart [by default, bar chart is inserted] select the dropdown and insert a "**Histogram**".

The histogram looks like:



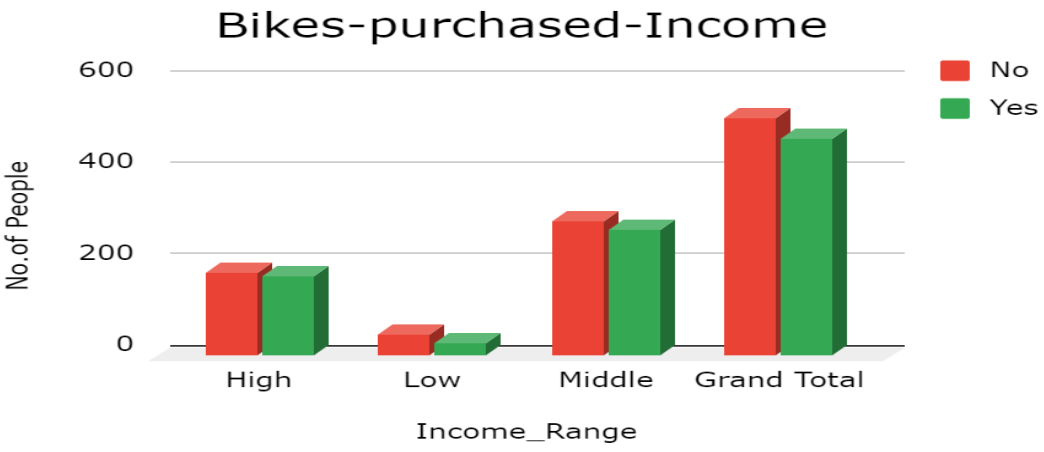
Step-2: Create a separate column named "income range" dividing the incomes into three ranges [low, middle, high] using "IF and AND functions". The formula looks like:

```
=IF(AND(K2>=10000,K2<20000),"Low",IF(AND(K2>=20000,K2<70000),"Middle",IF(AND(K2>=70000,K2<200000),"
High","Null")))
```

Step-3: Create a pivot table having columns with "**Income Range**" and rows with "**Purchased Bike**", values with "**Income Range**", Filter Null values in "**Income Range**".

Step-4: Select pivot table, select insert option and insert a chart. By default, "**Bar plot**" is inserted.

Step-5: Customize the plot by changing **colours, legend, serie**s etc to understand the plot easily by the user.



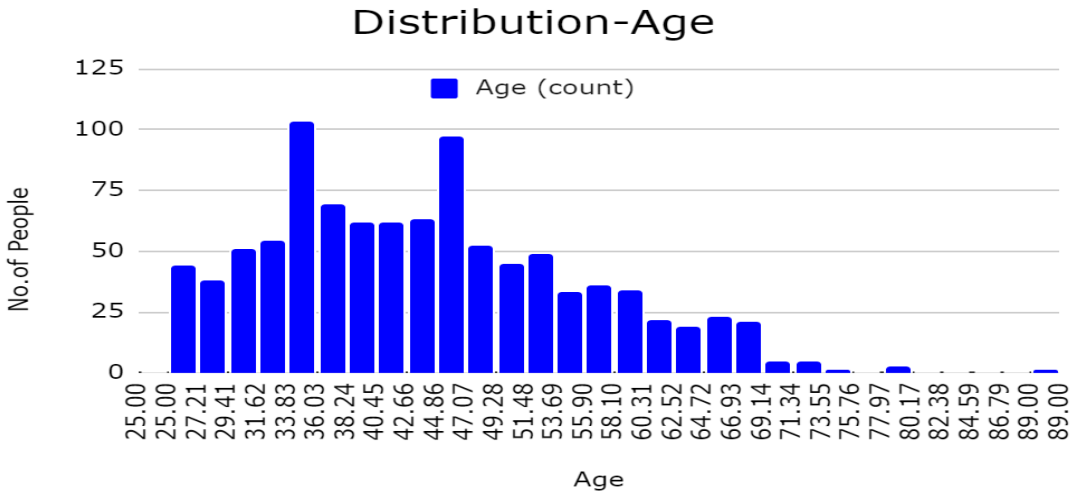
- Conclusion: Middle Income [20000-70000] people are purchasing bikes more than Low [<20000] and High [>70000] income People.
- In our Distribution Histogram we can observe that Middle income people are more than Low- and high-income people.

Histogram (Age):

Question: Create a histogram to understand the age distribution of bike buyers. Are certain age groups more inclined to purchase bikes?

Step-1: Select the column "**Age**" and insert a Chart [by default, bar chart is inserted] select the dropdown and insert a "**Histogram**".

The histogram looks like:



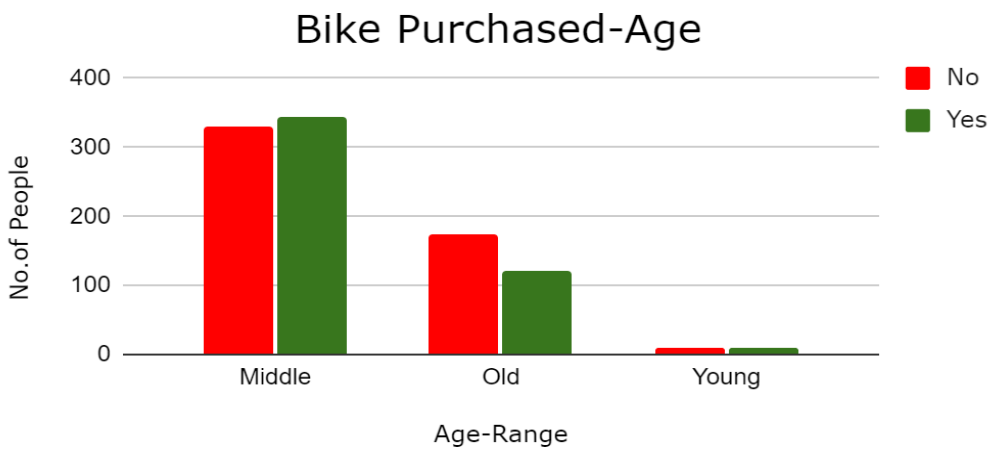
Step-2: Create a separate column named "**Income Range**" dividing the incomes into three ranges [Young, middle, Old] using "IF and AND functions". The formula looks like:

=IF(AND(L3>=18,L3<27),"Young",IF(AND(L3>=27,L3<50),"Middle",IF(AND(L3>=50,L3<75),"Old","Null")))

Step-3: Create a pivot table having columns with "**Age Range**" and rows with "**Purchased Bike**", values with "**Age Range**", Filter Null values in "**Age Range**".

Step-4: Select pivot table, select insert option and insert a chart. By default, "**Bar plot**" is inserted.

Step-5: Customize the plot by changing **colours, legend, series** etc to understand the plot easily by the user.



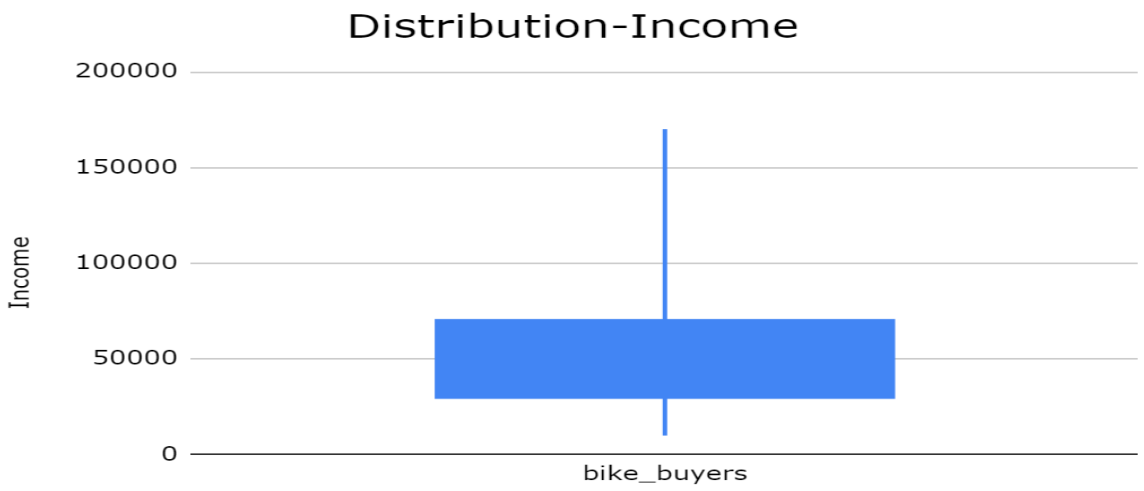
- Conclusion: Middle aged people [27-50] are purchasing bikes more than Young [18-27] and Old [>50] aged People. Age groups are inclined to purchase bikes.

Box plot (Income):

Question: Identify outliers in the income distribution of bike buyers. Are there any extreme income values, and how might they impact purchasing behaviour?

Step-1: Find Minimum, Maximum, first Quartile, Third Quartile, IQR [10000,170000,30000,70000,40000] respectively.

Step-2: Select these values and insert chart. By default, "Column Chart" is inserted. Change the chart to "Candlestick chart".



Step-3: Find “Lower Fence” and “Upper Fence” as “-30000” and “130000” respectively.

The values less than Lower Fence and greater than Upper Fence are “Outliers”.

- ❖ Conclusion: There are outliers in the data and they are marked as “Blue” in the dataset.
- ❖ There are extreme income values [>130000] effecting the target variable.

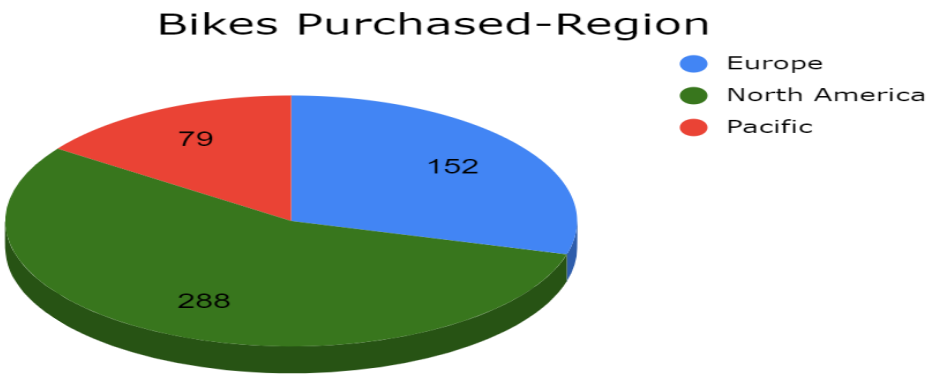
Pie Chart (Region):

Question: Represent the distribution of bike purchases by region using a pie chart. Are there regions where bike purchases are notably higher?

Step-1: Create a pivot table having columns with “**Region**” and rows with “**Purchased Bike**”, values with “**Region**”, Filter Null values in “**Region**”.

Step-2: Select pivot table, select insert option and insert a chart. By default, “**Bar plot**” is inserted change it to “**Pie Chart**”.

Step-3: Customize the plot by changing **colours, legend, serie**s etc to understand the plot easily by the user.



- Conclusion: From the pie chart we can observe that "**North America**" people are purchasing bikes more than "Europe" and "Pacific" regions.

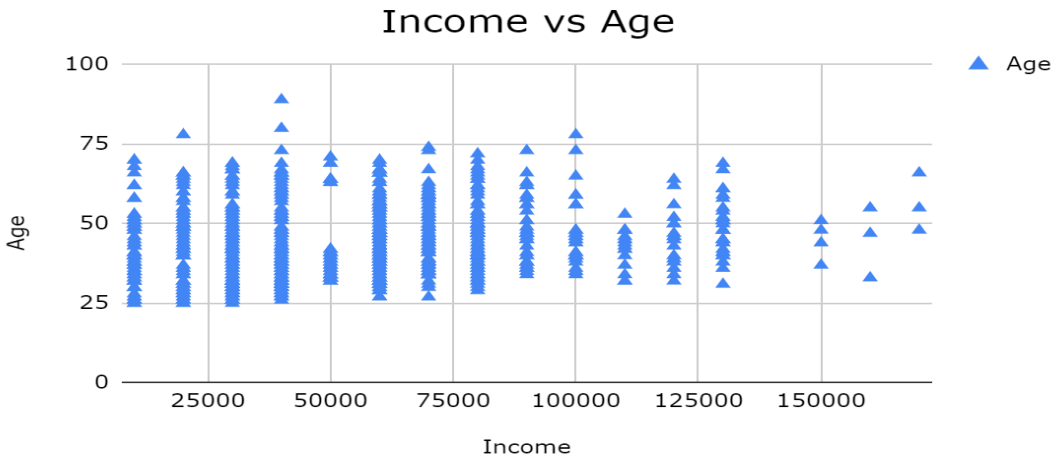
Scatter Plot (Income vs Age):

Question: Create a scatter plot to investigate the relationship between income and age. Do individuals with higher incomes tend to be in specific age groups?

Step-1: Select "Income" and "Age" by holding shift button.

Step-2: Insert chart. By default, "Column Chart" is inserted. Change it to "Scatter plot".

Step-3: Customize the plot by changing **colours, legend, serie**s etc to understand the plot easily by the user.



- Conclusion: From the scatter plot we can observe that middle and old aged people are having high income than the young people.

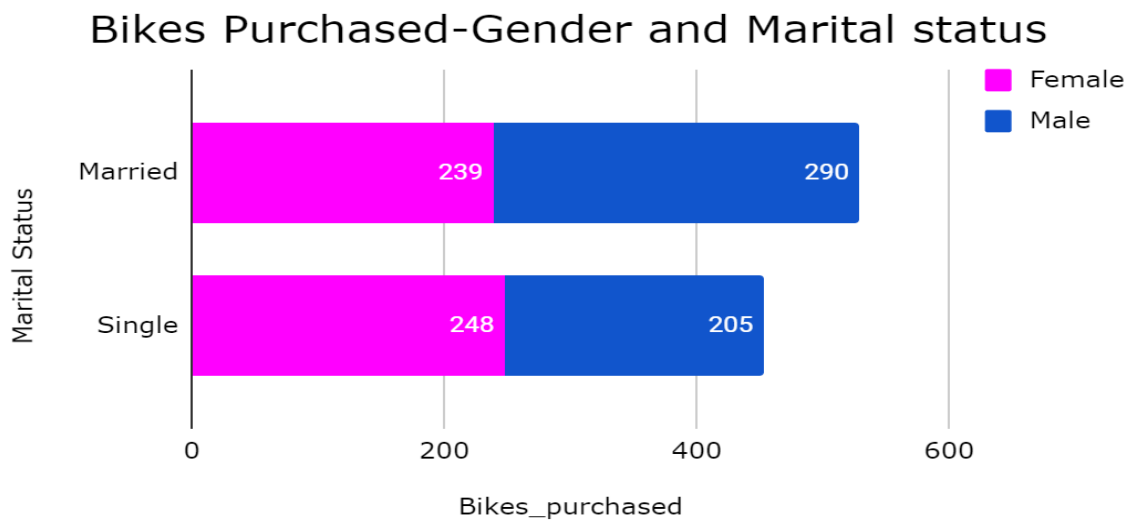
Stacked bar Chart (Marital Status and Gender):

Question: How does the distribution of bike purchases differ when considering both marital status and gender simultaneously? Are there notable patterns?

Step-1: Create a pivot table having columns with "**Marital Status**" and rows with "**Gender**", values with "**Purchased Bike**", Filter Null values in "**Marital status and Gender**".

Step-2: select the pivot table and insert a chart. By default, "Column Chart" is inserted. Change it to "Stacked Bar Chart".

Step-3: Customize the plot by changing **colours, legend, serie**s etc to understand the plot easily by the user.



- Conclusion: from the chart we can observe that "Single Women" and "Married Boys" are purchasing bikes more than the rest.

Correlation Heatmap (Numeric variables):

Question: Use a heatmap to visualize the correlation matrix between numeric variables. What variables show a strong correlation, and how might this influence purchasing behaviour?

Step-1: In our Dataset we are having only "Three Numeric Variables" those are Age, Income, Children.

Step-2: Select them and create and create a table with same three rows and columns [with the numeric variables].

Step-3: Same row and column variables have value 1.

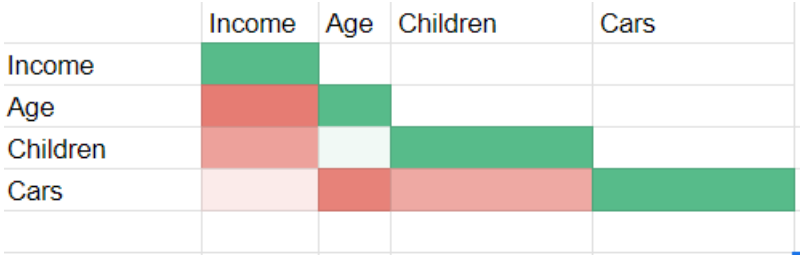
Step-4: Remaining Values are calculated by using the function “CORREL()”.

Step-5: After getting all Values the round them to two floating point integers.

Step-6: Delete “Grid Lines” for the Sheet and select “More Formats” and select “Custom Number Format” and type “;;;” in the tab and apply it to the selected cells.

Step-7: Select “Conditional Formatting” and go to “Color Scale” and select a color relevant to the data.

The final Heatmap is:



- Conclusion: From the Heatmap we can say that the “Age” and “Children” variables are having strong correlation that the other correlations.