Bike Buyers Data Analysis

Final Report

Jaydeep P. (jp157), Bhavya M. (brmistry)

INFO I-590 Data Visualization

Fall 2022

Table of Content:

Abstract

This project depicts people who purchased bicycles based on their geographical location, occupation, gender, age, commute distance, and people with vehicles. The visualizations would help the seller better understand the customer and provide them with the necessary bike. One of the major issues that the industry faces is that they are unaware of the trend of bike sales based on location and customer type, which can be solved by analyzing bike buyers' data sets for different locations. We have chosen three locations in our visualization: North America, the Pacific, and Europe.

Introduction

Background

In 2021, the market for bicycles was estimated to be worth USD 78.33 billion. The market is anticipated to expand at a 6.5% CAGR from 2022 to 2029, rising from USD 82.50 billion in 2022 to USD 127.83 billion in 2029. Bicycle demand has exceeded expectations in all locations compared to pre-pandemic levels because of the unprecedented and overwhelming COVID-19 pandemic. According to the research in [1], the global market grew by 48.6% in 2020 compared to 2019.

Americans continue to rely heavily on automobiles to get to and from work. According to Statista's Global Consumer Survey, 76 percent of American commuters drive themselves to work, making it the most common method of transportation.

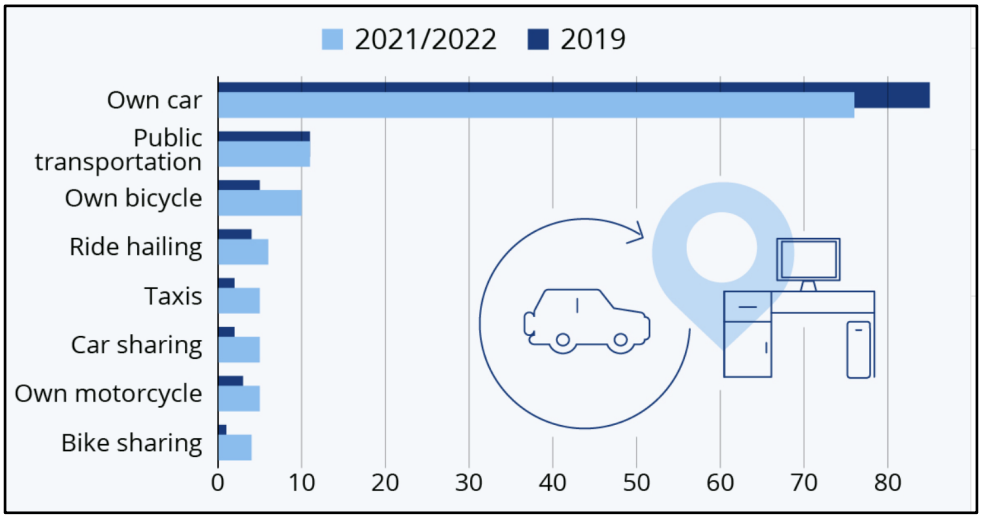

Figure 1: Based on multi-pick surveys of 3,338 (2019) and 5,649 (2021/2022) https://cdn.statcdn.com/Infographic/images/normal/18208.jpeg

U.S. adults who commute to work, A median of one-third (35%) of people in the 44 nations we examined say they have a working car at home. Americans rank towards the top, with 88% claiming to own a car, and are on pace with Italians (89%). A median of 79% of respondents in the seven European Union countries polled possesses a car. Other industrialized economies, such as South Korea and Japan, have high rates of car ownership.

A median of 42% of people in the 44 nations said they have a working bicycle at home. However, only roughly half of Americans (53%) possess a bicycle. Bicycle ownership is more prevalent in Asia, the EU, and the United States than in Africa and the Middle East. It is also more common in developed economies than in developing economies [3]. This was the major issue faced by bike sellers, as despite having great infrastructure and better commutes, how bike sellers could expand their business and succeed in targeting the customers.

Motivation

Countries such as the United States have maintained Bike lanes and road infrastructure, enabling bicyclists to ride at their preferred speed without interference from prevailing traffic conditions and facilitating predictable behavior and movements between bicyclists and motorists. Though bicycles are ideal transit for citizens still they prefer to commute using cars, buses, and trains.

It was encouraging to see how many people preferred commuting using bicycles to cut traffic, and traveling according to their own time, also contributing to the environment by reducing carbon emissions and maintaining their health.

Despite having great commutes for bikes, there was a comparatively small number of people driving them. After analysis, we came to know that many peoples enquired about bikes but, shop owners sold the same type of bikes and all of a similar range, also the location where the store was located was not quite accessible to certain groups of people, as the many people avoided using bike because they need maintenance, and the location was quite far off. The best way shop owners could determine their user base was by knowing some information about the buyers.

Objective

The objective was to learn about the increase or decrease trend in the use of bikes by people. One of the primary objectives was to help the shop owners to expand their business, which could have been possible if they knew their customers right, thus improving customer profitability, improving brand loyalty and Customer retention Helping them was eventually helping the environment and people. As more people use bikes their body remains fit and has fewer health complications, also carbon emission, and low pollution thus reducing global warming.

This piqued our interest, and we decided to investigate the data pertaining to the people who bought bikes. There are data visualizations related to bike sales, but they cannot assure the reason for the person purchasing the bike. The visualizations are somewhat confusing; thus, it would have been clearer and crisper for the type of buyer for the bike.

Related Work

The Visualizations for sales of the bike are available on most of the websites and the graphs given are not that intuitive to understand the customer. For example, the image provided below focuses on the types of bikes that were sold but there is no information related to the type of person buying those bikes, and the visualization is monotonous not getting us to study sales of the bikes [2].

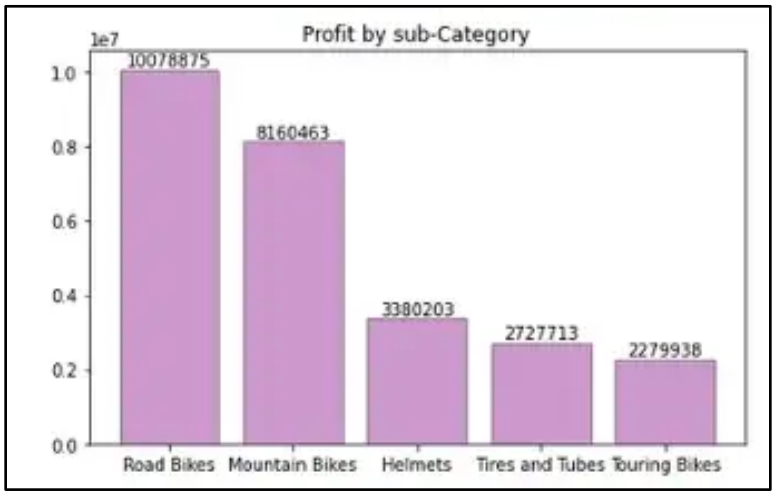

Figure 2: Most profitable category of the bike sold from the given categories

https://miro.medium.com/max/828/1*ONIxmodInbGG25u0LD9Mbg.webp

The visualizations present over the internet are not interactive over the region. We have tried to segregate the graph based on the location of the customer to get a better perspective regarding the regional aspects of people. The Data is solely based on the sales of the bikes and the revenue, but it can be improved if the bike shops also focus on the customer type, and they can get the probability of the person buying the bike or not when they visit the shop.

The visualizations available for the bike sales are very much sales-centric and not focused on the customer if the bike seller gets aware of the customer's specifications such as their income, commute distance, and multiple other factors which would influence the customer to purchase the bike it would be beneficial for the sellers to improve the sales of the bike, as more number of bikes used for traveling, less carbon emission thus reducing global warming.

Existing Visualizations

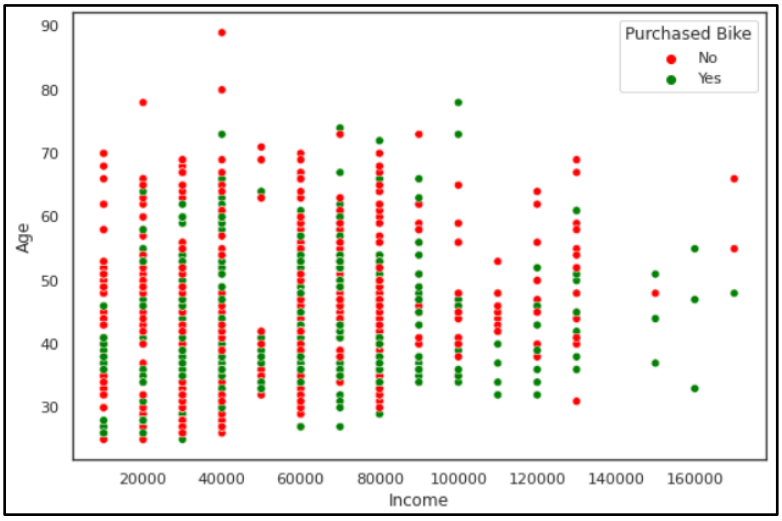

Figure 3: Scatter plot for Income vs Age who purchase bike or not

The above diagram was taken from the Kaggle notebook where we are not properly able to identify the points individually from the cluster which leads to occlusion of the points leading to the incorrect data representation for the dataset. It would have been clearer if the size of the marker was smaller but that doesn't guarantee occlusion-free visualization. Also, the color choice of the markers is not colorblind friendly, since the colors used are a combination of red and green. We are improving the case by using jittered scatter plots and even there is no segregation for the region-specific public which we have improved in our visualization by providing the visualization interactive based on regions. So, it gets more articulate to understand the potential buyer of the bike.

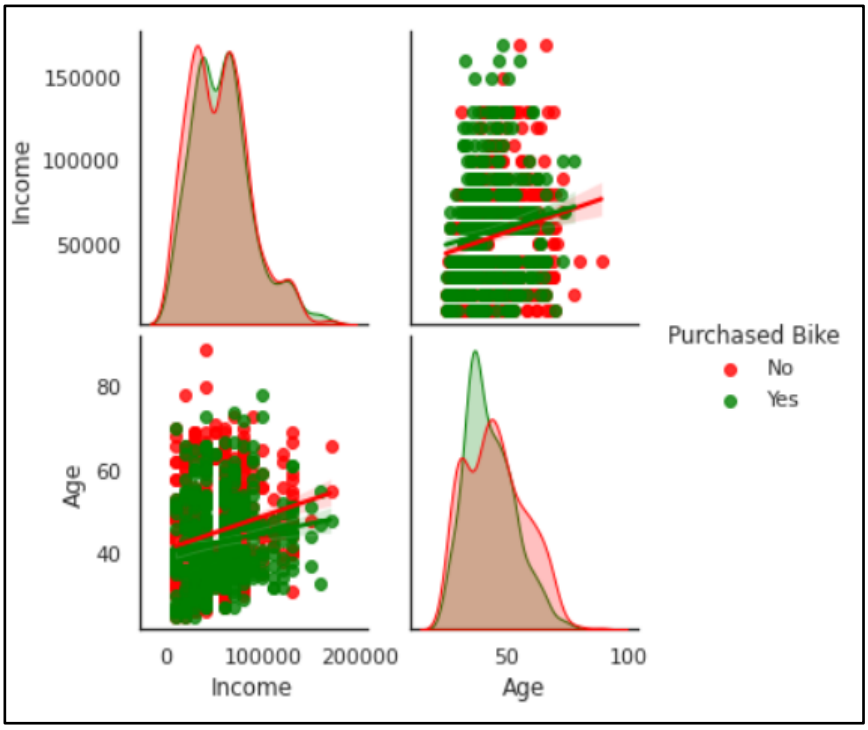

Figure 4: Pair plot for Income and Age parameters

In the above diagram as we can see that the color chosen is not a good choice because the diagram is not for color-blind friendly people like Figure 3 and the size of the points are very big which in turn leads to unclear outcomes. The income could be more informative if the visualization also provided the kind of occupation that belongs to the income using interactive visualization techniques.

Bike Buying Prediction for adventure works [4]

The project mentioned in [4] uses classification and makes predictions on whether a person can buy a bike or not for adventure tracks. They considered the bike-sharing system and used a variety of prediction models to determine the optimum algorithm with the highest accuracy level. The production and accuracy have both been compared to the model's side by side. This technique incorporates that to assess the likelihood of producing a workable system.

The problem with bike stations, which arises when a customer visits a bike station and discovers that there is nowhere to return the bike, may be resolved using predictions. The bike redistribution may be prepared for if it is known how many bikes would be required each day. Additionally, the problem of theft and system abuse cannot be solved. The random forest model produced accurate findings. The managers must come up with creative ideas to keep the motorcycles secure. The subject of bicycle riders' safety is examined; in order to protect users, bicycles should always be mounted.

Data and Method

Data Analysis

We have done an analysis of bike buyers and see which kind of person will buy a bike or not. In order to check whether people will buy a bike depending on their needs, we have used the data set from Kaggle. We first examined and queried the data before doing an evaluation. We Interpreted the data and cleaned the dataset for the visualizations. We have the following attributes in our dataset:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 13 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   ID                1000 non-null   int64
 1   Marital Status    993 non-null    object
 2   Gender            989 non-null    object
 3   Income            994 non-null    float64
 4   Children          992 non-null    float64
 5   Education         1000 non-null   object
 6   Occupation        1000 non-null   object
 7   Home Owner        996 non-null    object
 8   Cars              991 non-null    float64
 9   Commute Distance  1000 non-null   object
 10  Region            1000 non-null   object
 11  Age               992 non-null    float64
 12  Purchased Bike    1000 non-null   object
dtypes: float64(4), int64(1), object(8)
memory usage: 101.7+ KB
```

Figure 5: Database Schema

There are 13 variables in this dataset:

1. ID: This is the integer datatype used to uniquely identify a row.
2. Marital Status: It is a non-numeric data type that shows the status of the person being married or single.
3. Gender: A non-numeric data type that states whether the buyer is Male or Female.
4. Income: It is a numeric data type that shows the Income of the person.
5. Children: The number of children. The children column type is non-numeric.
6. Education: Buyers' educational status like bachelor's, Graduate Degree, High School, Partial College, or Partial High School. It is a non-numeric data type.
7. Occupation: The buyer's job from Clerical, Management, Manual, Professional, or Skilled Manual. It is a non-numeric data type.
8. Homeowner: Indicates whether the buyer has his/her own house or not. The data type of the Homeowner column is non-numeric.
9. Cars: Number of cars the buyer has, and it has data type as integer.
10. Commute Distance: Commute distance shows the distance between the buyer's house and the bike shop. There is 0-1 Miles, 1-2 Miles 10+ Miles, 2-5 Miles, and 5-10 Miles ranges in the column. It is of non-numeric data type.
11. Region: There are three places from where buyers are there such as Europe, North America, or the Pacific. This region attribute has a non-numeric data type.
12. Age: It shows the age of the buyer. It has a data type of integer.
13. Purchased Bike: Whether the buyer has purchased a bike or not it is of character data type.

Here is a peek of the dataset we have used,

	ID	Marital Status	Gender	Income	Children	Education	Occupation	Home Owner	Cars	Commute Distance	Region	Age	Purchased Bike
0	12496	Married	Female	40000.0	1.0	Bachelors	Skilled Manual	Yes	0.0	0-1 Miles	Europe	42.0	No
1	24107	Married	Male	30000.0	3.0	Partial College	Clerical	Yes	1.0	0-1 Miles	Europe	43.0	No
2	14177	Married	Male	80000.0	5.0	Partial College	Professional	No	2.0	2-5 Miles	Europe	60.0	No
3	24381	Single	Male	70000.0	0.0	Bachelors	Professional	Yes	1.0	5-10 Miles	Pacific	41.0	Yes
4	25597	Single	Male	30000.0	0.0	Bachelors	Clerical	No	0.0	0-1 Miles	Europe	36.0	Yes

Figure 6: Database Five Rows

Libraries/Modules Used

Pandas– It was used to analyze and manipulate the dataset.

NumPy- NumPy was used for some mathematical operations.

Matplotlib- To improve the GUI and plot the visualizations.

Seaborn- Seaborn is used for plotting some complicated visualizations.

Ipywidgets- Ipywidgets are used to make the jupyter notebook interactive.

Process

In this project, we are going to use the ‘bike_buyers.csv’ file. After importing the data, we check if there are any null values, if any we make changes to the dataset using transformations like cleaning, and replacing, them with better options. Let’s examine the data first.

```
df = pd.read_csv("bike_buyers.csv")
```

```
df.isna().sum()
```

```
ID                   0
Marital Status       7
Gender              11
Income               6
Children             8
Education            0
Occupation           0
Home Owner           4
Cars                 9
Commute Distance     0
Region               0
Age                  8
Purchased Bike       0
dtype: int64
```

Figure 6: Database Five Rows

Data Columns Marital Status, Gender, Income, Children, Homeowner, Cars, and Age have NA values. To fill in the values, we fill the data with transformations like cleaning, replacing, and adding calculated columns. These are the steps that clean and transform our data set:

```
def Fill_Missing_Value(df):
    ind=0
    med = 0
    for i in df.dtypes:
        if i == 'float64' or i == 'int64':
            column_name=df.columns[ind]
            column_value=df[column_name][0:]
            med= column_value.median()
            df[column_name] = df[column_name].fillna(round(med,1))
        if i == 'object'or i=='o':
            column_name=df.columns[ind]
            column_value=df[column_name][0:]
            mode= column_value.mode().values[0]
            df[column_name] = df[column_name].fillna(mode,inplace=False)
        ind+=1
    return df
df=Fill_Missing_Value(df)
df.isna().sum()
```

```
ID                  0
Marital Status      0
Gender              0
Income              0
Children            0
Education           0
Occupation          0
Home Owner          0
Cars                0
Commute Distance    0
Region              0
Age                 0
Purchased Bike      0
dtype: int64
```

Figure 7: Filling Missing Values

We fill the columns (Income, Children, Cars, Age) having Integer and Float values with median, and columns (Marital Status, Homeowner,) having non-numeric values are filled using mode.

In order to find the correlation between columns, we draw a heatmap to understand which columns are useful and which columns can be dropped and aren't useful in analysis and predicting whether users will purchase a bike.

Figure 8: Correlations between the columns

According to figure 8, column ID has the least relation with other columns compared to others. Thus, if we drop column ID, data will become more consistent.

Next, in order to check which columns having numeric values contain any outliers, we can infer from the below box plots that the Income attribute has three outliers in it which lie between 150000 to 170000. In the Cars attribute, we found that there existed one person in the dataset who was having 4 cars. For Children's columns, there are no outliers and for Age there were 3 outliers, belonging in the range of 78-90.

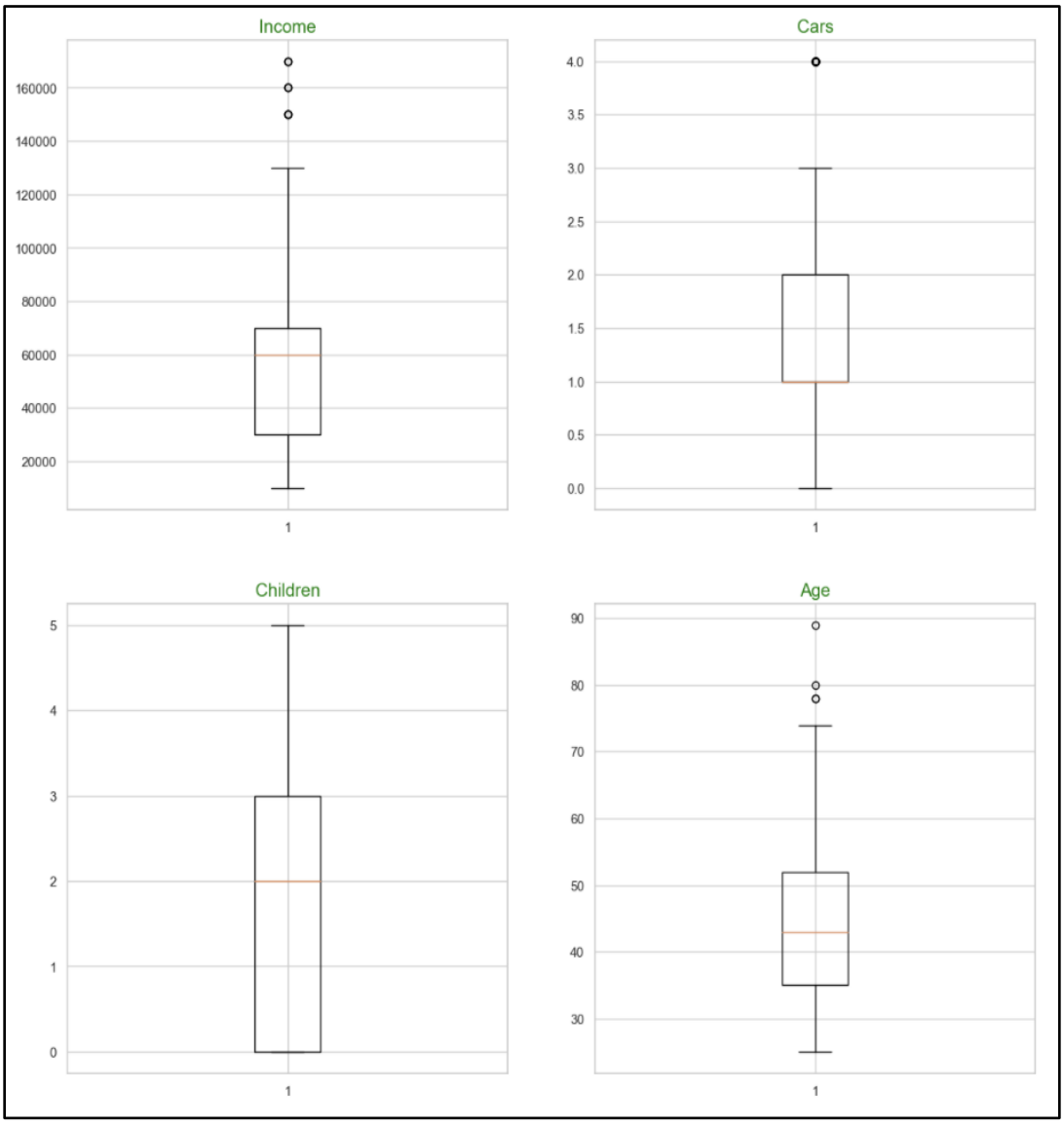

Figure 9: Box Plot to show outliers

To understand the effective variables in the data set we plotted the pair plot matrix of the data and made hue according to the region to understand the data set region-wise. The attributes of Income, age, and cars were the important features that we wanted to use in the visualizations.

Figure 10: Pair Plot to see relationships between two variables

Failed Experiments

While visualizing, many things matter such as the choice of columns, axes of the plot, color, marker size, and clear insights. In order to depict clear visualizations we should avoid representing columns on the scale, which doesn't match and makes the visualization unclear and creates more confusion.

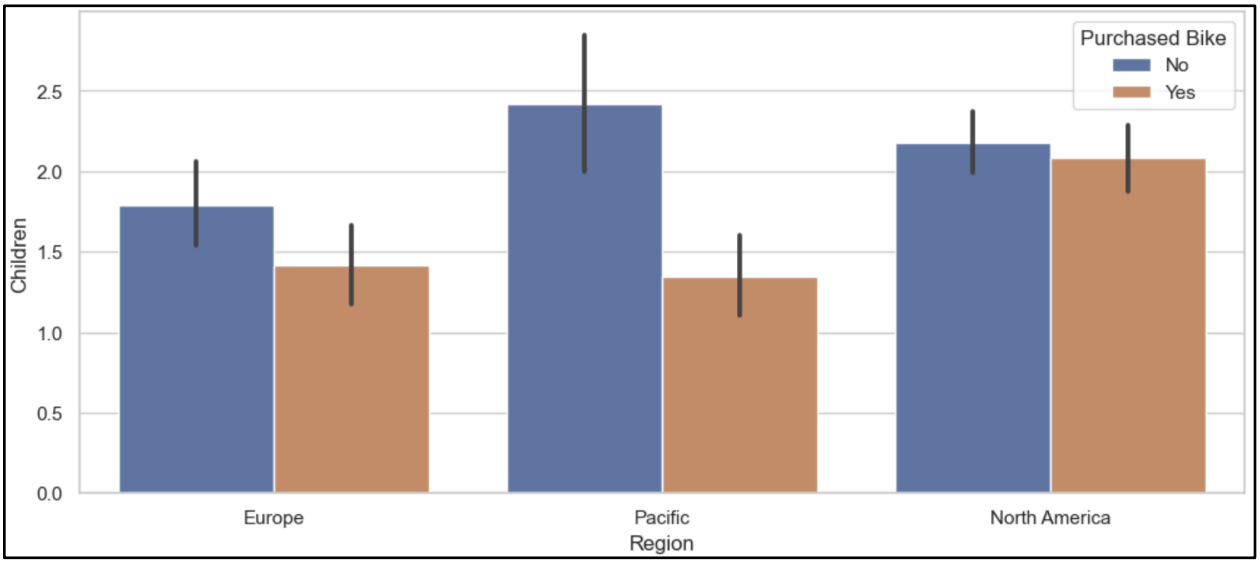

Figure 11: Bar Plot showing unclear values on the Y axis

In this figure above, the columns Children plotted on the Y-axis, and Region plotted on the X-axis with hue as Bike Purchased. So, the issue in the plot is that, when carefully observed, many people who purchased bikes residing in Europe have 1.8 children, which is not possible. Thus, we solved the issue by replacing the plot with axes as depicted in the figure below. Where currently it shows

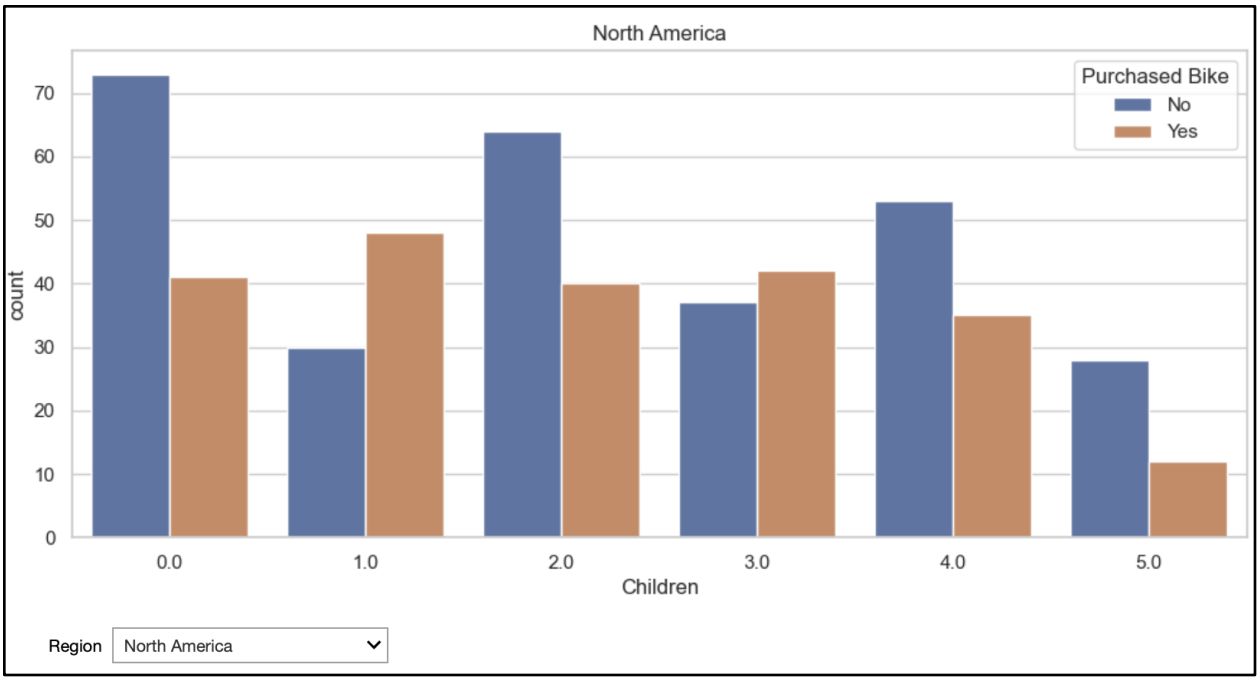

Figure 11: Counting people buying bikes with children and regions correlating

Similarly, in the violin plot below the Cars are plotted on Y-axis and Region plotted on X-axis with the purchased bike. The starting limit is -1 which is not the correct value for column Cars. So, we can correct the plot below by limiting the Y-axis or starting the Y-axis with 0.

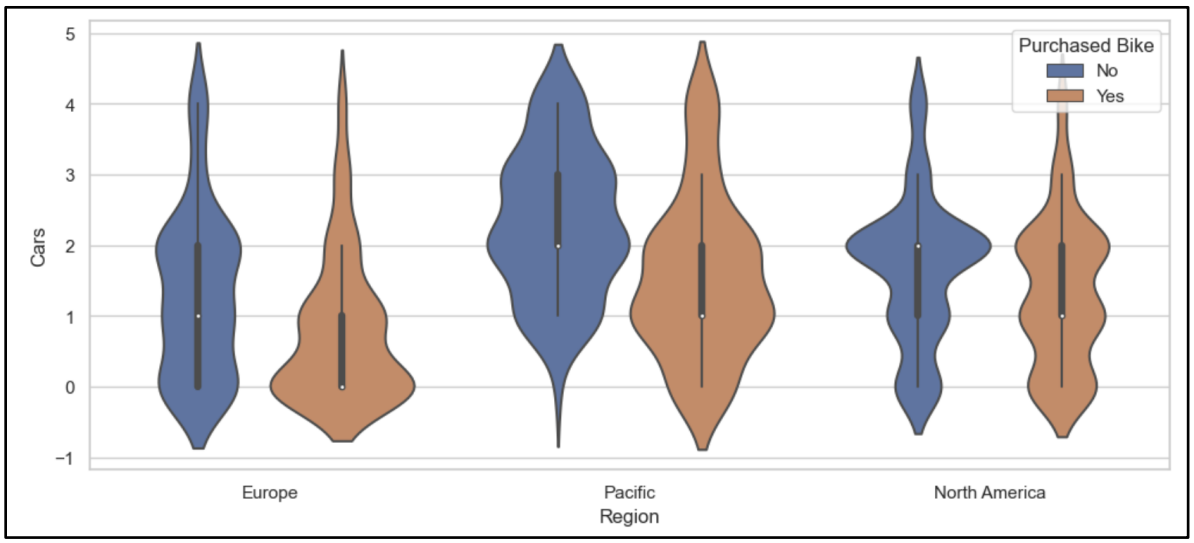

Figure 12: Violin plot not starting Y-axis from 0

Results and insights

Visualizations and their explanation

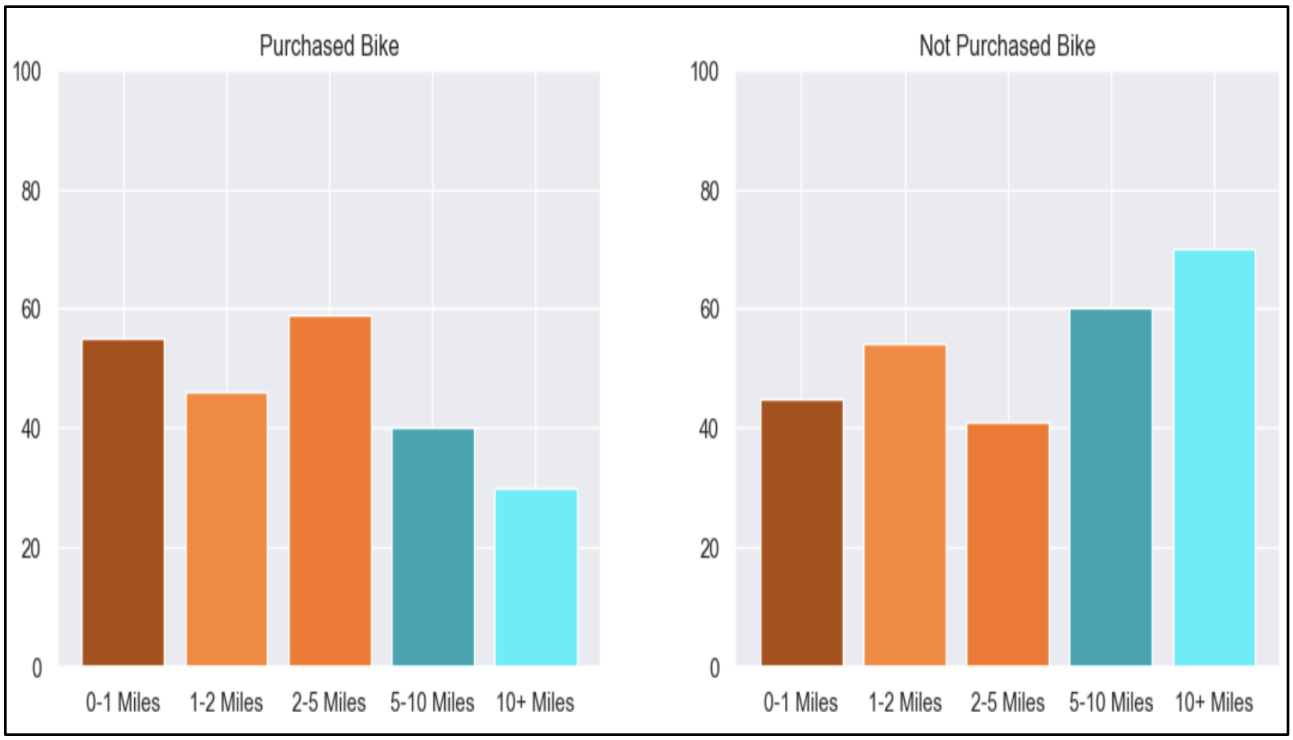

Figure 13: Bar plot to visualize

Bar plot of the commute distance vs the percentage of people buying the bike or not, we can conclude from the above plot that most people leaving nearby in the range of 0-1 and 2-5 miles of the shop are more likely to purchase the bike and commute distance more than 5 miles there are very fewer chances that the person buys the bike.

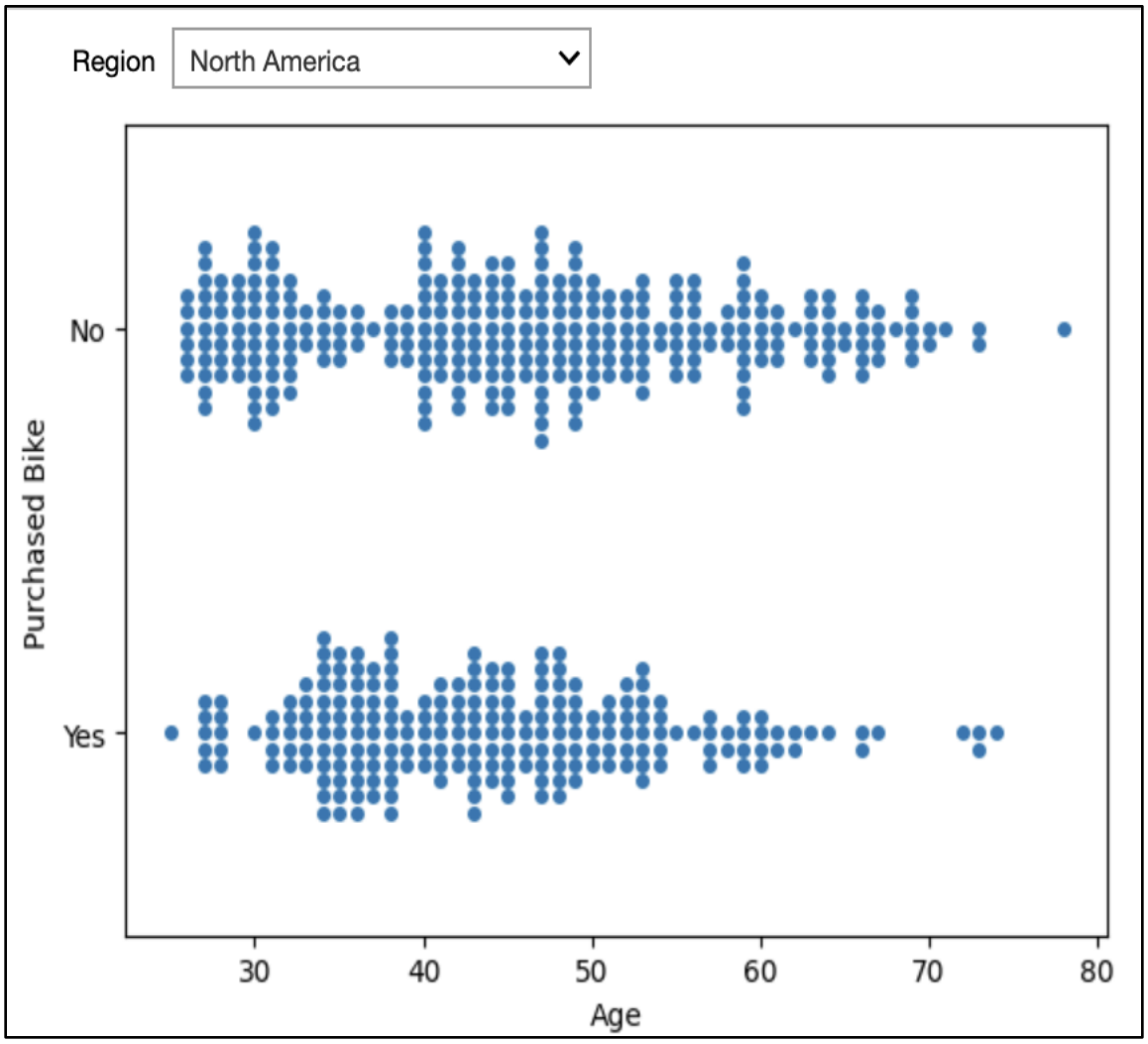

Figure 14: Interactive jittered scatterplot (Beeswarm Plot)

The jittered scatterplot (beeswarm) plot for the age of the person and the bike purchased gives us the insight that people might be more health conscious between the age of 30-40 years as there are more bikes purchased. After the age of 40, there are fewer people tend to buy a bike.

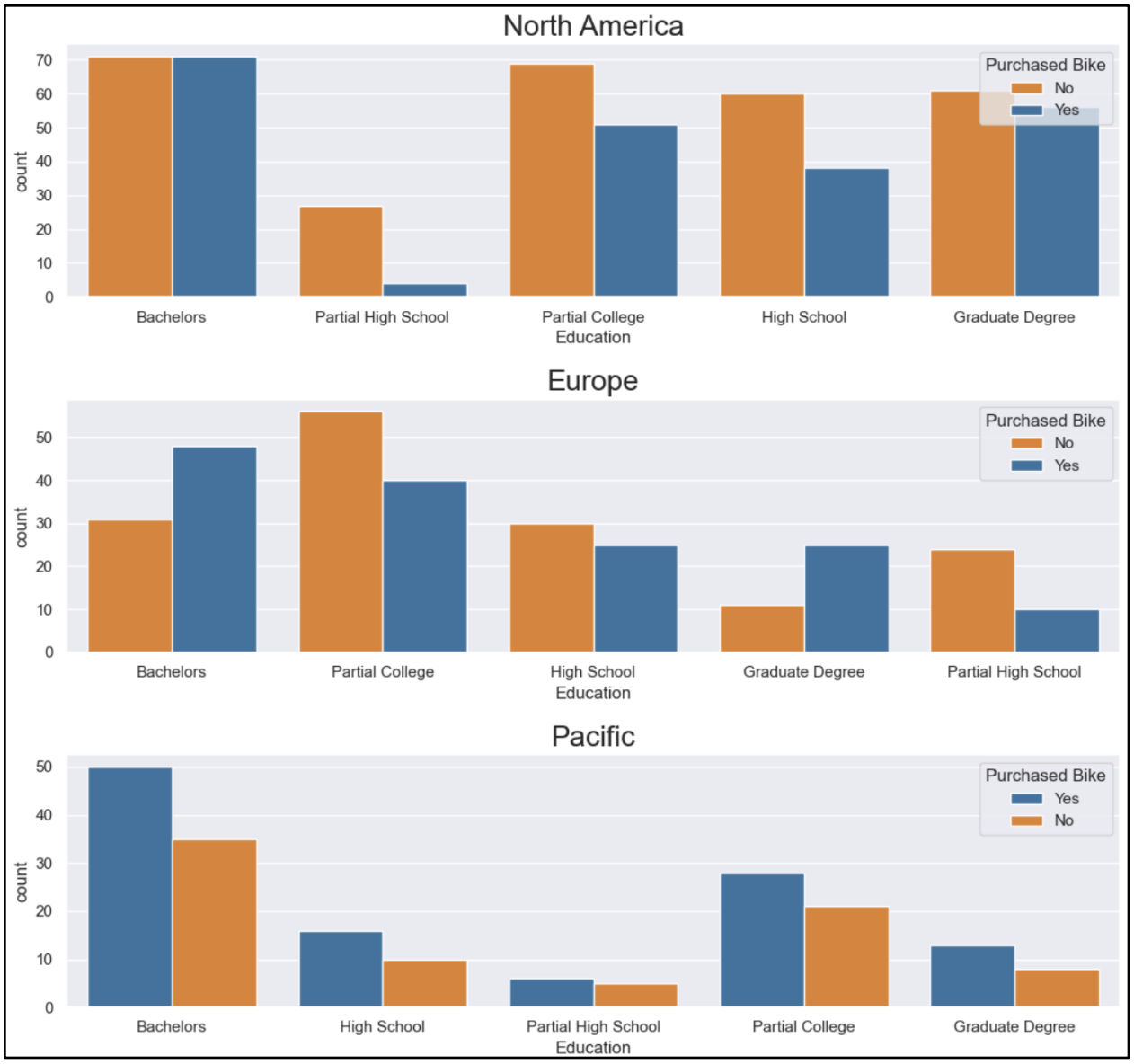

Figure 15: Counting People owing bike purchase region-wise with Education

We segregated the data into 3 parts according to the regions- North America, Europe, Pacific and plotted the bar graph for the occupation of the people and whether they have purchased a bike or not. And we can infer from the plot for North America that there were an equal number of bachelors who bought bikes and there were very few people around 40, who did partial high school and were owning a bike.

For Europe, the maximum number of people owning a bike had completed their bachelor's, and the least were from a partial high school of around 10 people. In the Pacific region also, the maximum bikes were owned by bachelors and the least bikes were there for partial high school education people.

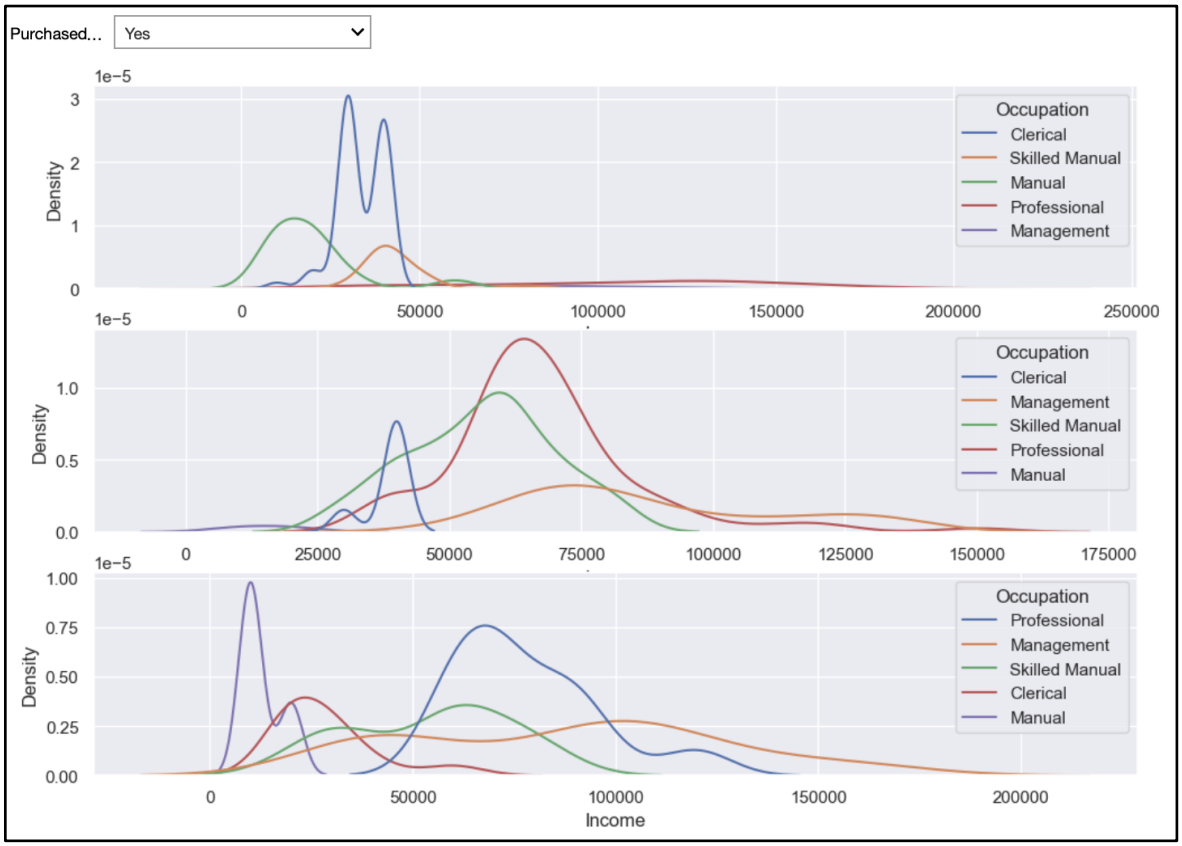

Figure 16: Correlating Education with bike purchase region wise

The above KDE plot is for different occupations and their earnings with bikes purchased as interacted. According to the plot, North America indicates that there are maximum people with clerical occupations purchasing the bike followed by manual and skilled manual occupations, the least type of occupation who are not purchasing bikes are professionals and Management people.

For Europe, we can see that professional and skilled manuals earn around 50000 to 100000 purchasing bikes are maximum, whereas management occupation people are the least who bought bikes. In the Pacific region, most bikes were purchased by people having manual type of occupations also professionals earning from 50000 to 100000 have more bikes.

Figure 17: Histogram of House owned with marital status, region wise showing people purchase a bike

The above histogram with KDE distribution for people being married or single with whether they own a house or not gives us another insight that the people who are married and who own a house in North America have a maximum purchase of a bike and married who don't own house have the least number of bike in all the three regions.

The above plot can be clearly understood by comparing it with the plot below, depicting people who didn't purchase the bike.

Figure 18: Histogram of Houses owned with marital status, region wise showing people not buy a bike

Conclusion and future work

Visualizations answered multiple questions such as which gender purchases more bikes? What kind of profession of people with a particular range of income has a relationship with the sales of the bike? Does having a car impact the chances of a person buying the bike or not? And many more. We have concluded from the bar plot that the commute distance between 0-1 and 2-5 miles are more likely to purchase a bike. The age group of 30-40 purchases more bikes and after 40 there are very less people buying bikes. Students pursuing bachelor's from all the regions have more bikes as compared to other educational levels.

With the KDE plot, we concluded that in North America clerical, in Europe Professionals and pacific manual and professional occupations have the most bike purchases. Regarding marital status, we can conclude that North Americans who own a house and are married have a maximum number of bikes.

The effective way of visualizing the data of the bike buyers can give us multiple insights which we concluded after performing data visualization on the data set. In the future, this visualization can be helpful for bike sellers to increase sales, by knowing the customer and thus opening one or more stores in such locations where people living far from home can visit the stores for maintenance and other stuff. Also, online sellers can ask the customer for some demographic information, so that on that basis the cycles and cycle store can be recommended.

References:

1. Bicycle Market Size, Share & Value | Trends Analysis [2029]
2. Case Study: Bike Sales Analysis using Python. | by Reginald Charles | Medium
3. Car, bike, or motorcycle? Depends on where you live | Pew Research Center
4. https://www.kaggle.com/rahulsah06/bike-buying-prediction-for-adventure-works-cycles#AW BikeBuyer.csv
5. https://www.kaggle.com/code/r1azmahmud/analysis-and-prediction-of-bike-buyers-1000-data
6. https://valuationresources.com/Reports/SIC5941BicycleStores.htm
7. https://www.expertmarketresearch.com/reports/bicycle-market
8. https://www.digitalocean.com/community/tutorials/seaborn-kdeplot
9. https://medium.com/geekculture/python-seaborn-statistical-data-visualization-in-plot-graph-f149f7a27c6e