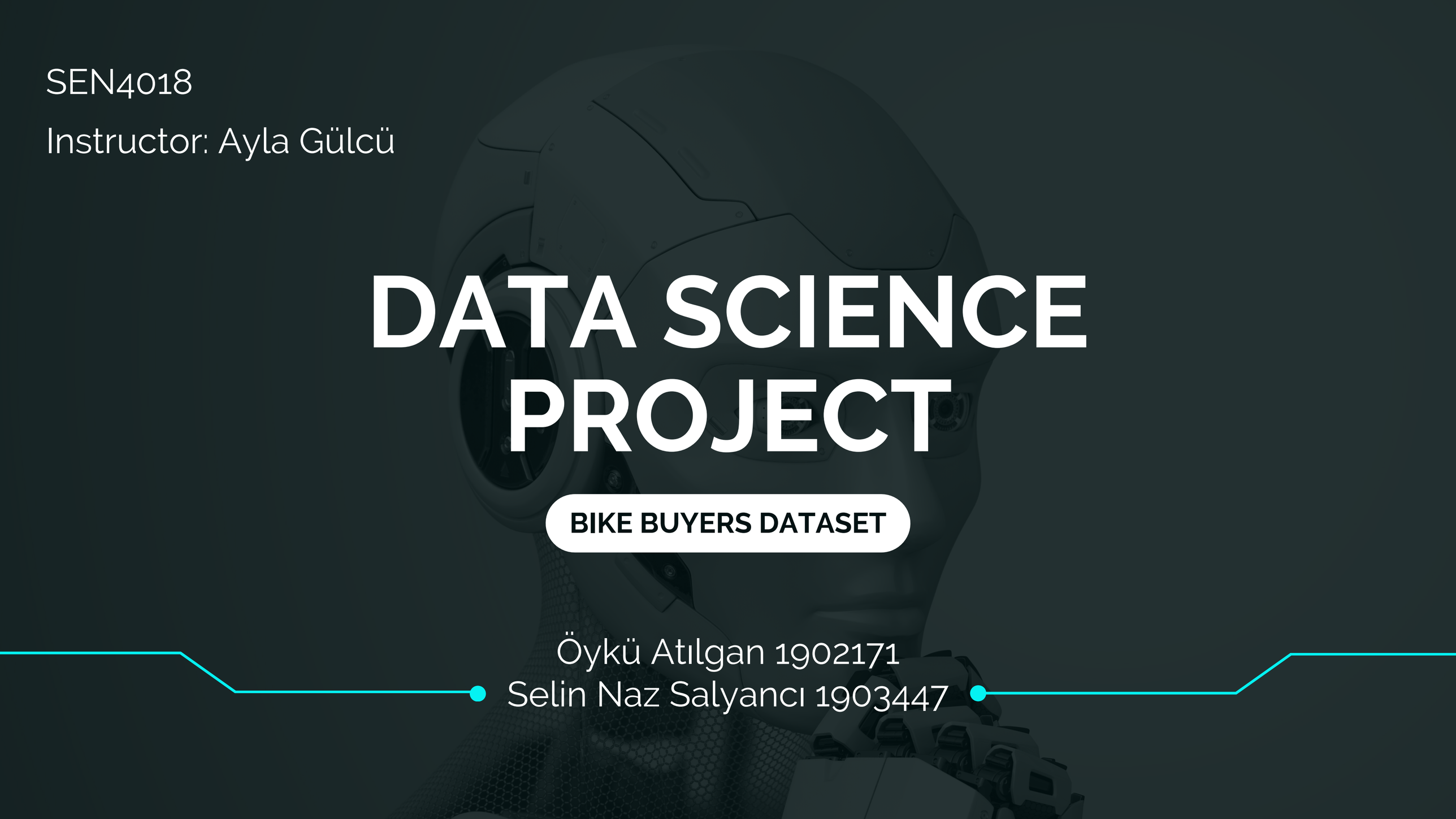
SEN4018

Instructor: Ayla Gülcü



DATA SCIENCE PROJECT

BIKE BUYERS DATASET

Öykü Atılgan 1902171
Selin Naz Salyancı 1903447

PART I

Brief summary of the dataset

The data set that we selected has 13 attributes and 1000 samples including missing/null values.
This set is created to represent the customers whether they purchased a bike or not. Also some informations about customers is provided in this dataset.

Aim Of the Study

The aim of this study is to implement a real world data science project with all steps included.

Major findings of the analysis

It was observed that the education level directly affected the bicycle purchase.

Contrary to education level, bicycle purchases decreased as income increased.

#Contrary to the above-mentioned results, the occuption level did not appear to have a complete effect on bike uptake.

According to age data, the middle age group generally has a high bicycle intake.

The rate of bike purchase in the Pacific region is higher than in other regions.

PART II

Attribute Name	Range	Attribute Type
customer_id	[11000,29400]	Nominal (numerical)
marital _status	[Married ,Single]	Nominal
gender	[Female,Male]	Symmetric Binary
income	[10000,170000]	Numeric(quantitative)
children_number	[0,5]	Numeric(quantitative)
education_status	[Bachelors, Partial College, High School,Graduate Degree,Partial High School]	Ordinal
occupation	[Professional, Skilled Manual, Clerical, Management, Manual]	Nominal
owns_home	[yes, no]	Symmetric Binary
car _number	[0,4]	Numeric(quantitative)
work_distance		nominal
location	[North America, Europe, Pacific]	Nominal
age	[25,89]	Numeric(quantitative)
purchased	[yes ,no]	Symmetric Binary

Description of the dataset

The data set represents bike buyed customers.
There are 13 variables in the data set;

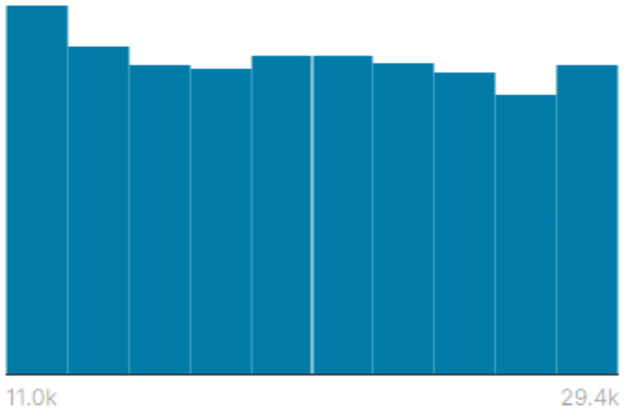
8 numerical

5 categorical

Basic Statistical Details

ID

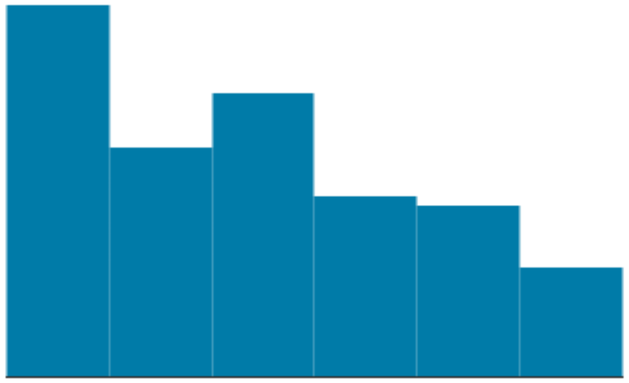
Customer ID



Valid	1000	100%
Mismatched	0	0%
Missing	0	0%
Mean	20k	
Std. Deviation	5.34k	
Quantiles	11.0k	Min
	15.3k	25%
	19.7k	50%
	24.5k	75%
	29.4k	Max

Children

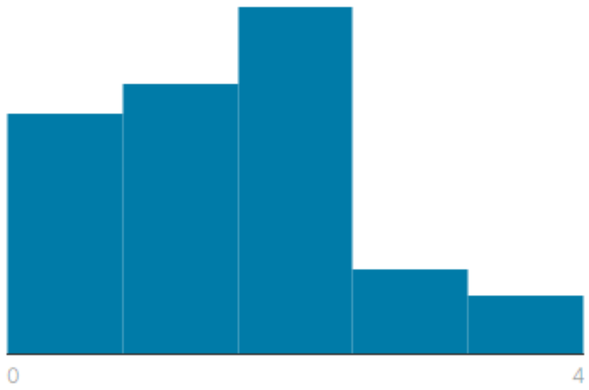
No. of children



Valid	992	99%
Mismatched	0	0%
Missing	8	1%
Mean	1.91	
Std. Deviation	1.63	
Quantiles	0	Min
	0	25%
	2	50%
	3	75%
	5	Max

Cars

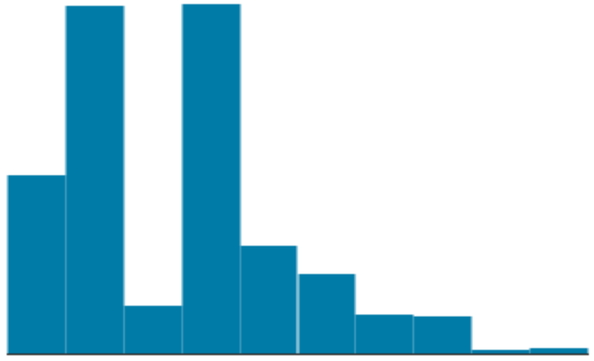
No. of cars owned



Valid	991	99%
Mismatched	0	0%
Missing	9	1%
Mean	1.46	
Std. Deviation	1.12	
Quantiles	0	Min
	1	25%
	1	50%
	2	75%
	4	Max

Income

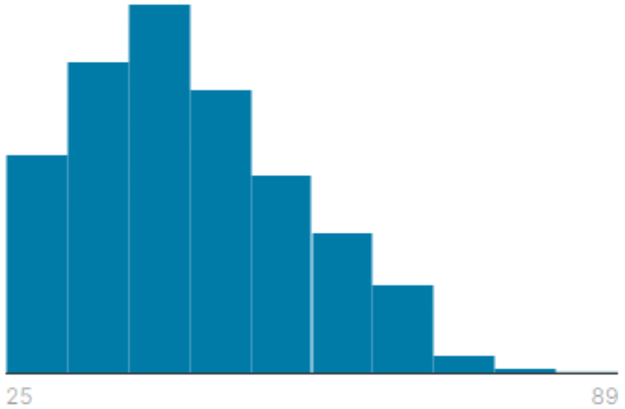
Annual Income



Valid	994	99%
Mismatched	0	0%
Missing	6	1%
Mean	56.3k	
Std. Deviation	31.1k	
Quantiles	10000	Min
	30.0k	25%
	60.0k	50%
	70.0k	75%
	170k	Max

Age

Age of customer



Valid	992	99%
Mismatched	0	0%
Missing	8	1%
Mean	44.2	
Std. Deviation	11.4	
Quantiles	25	Min
	35	25%
	43	50%
	52	75%
	89	Max

Step By Step Process

Step 01

Data Preprocessing

Step 02

Data Cleaning

Step 03

Data Integration

Step 04

Data Reduction

Step 05

Data Transformation and Data Discretization

PART III

Data Management Processes

**Does your data come from separate files? If so, how did
you merge them?**

Our data come from just one dataset file.

Did you apply any data transformation method?

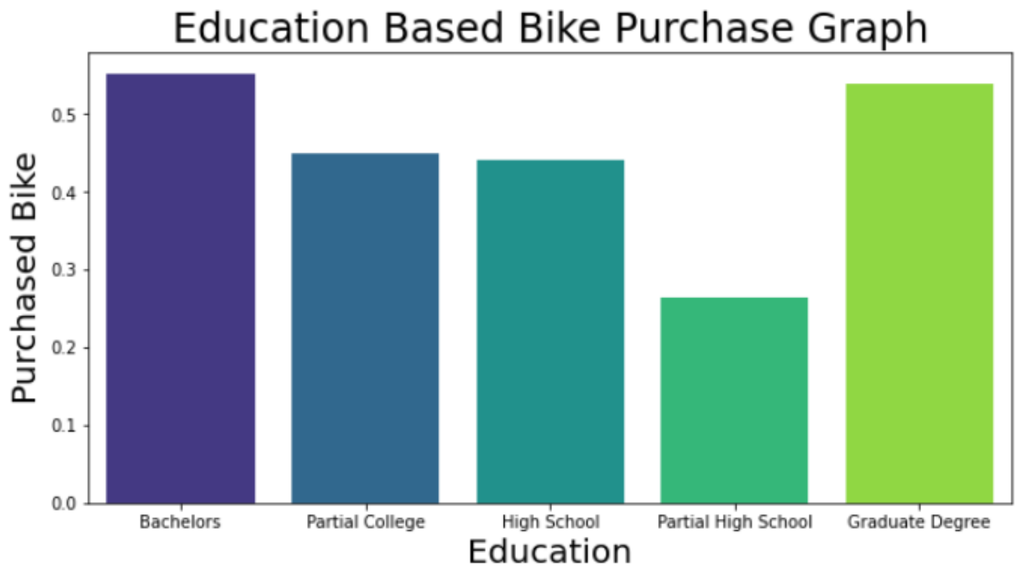
Yes, we changed “purchased bike” boolean value to integer data.

How did you handle missing values?

We handle missing values with two ways. Because some of the data was suitable for replacing values but some of them makes more sense when we use the mean value. First one was to replace null values as “No Data” with the usage of “fillna" function. The second way was to take mean as null vales.

PART IV - V

In [5]:
```
#create chart according to education
fig, bx=mplt.subplots(figsize=(10, 5))
visualA=sbrn.barplot( data = ds,x = 'Education', y = 'Purchased Bike', ci=0,palette="viridis",saturation=10)

#physical property adjustment of the graph
visualA.set(title="Education Based Bike Purchase Graph")
visualA.xaxis.get_label().set_fontsize(20)
visualA.yaxis.get_label().set_fontsize(20)
visualA.title.set_fontsize(24)
```



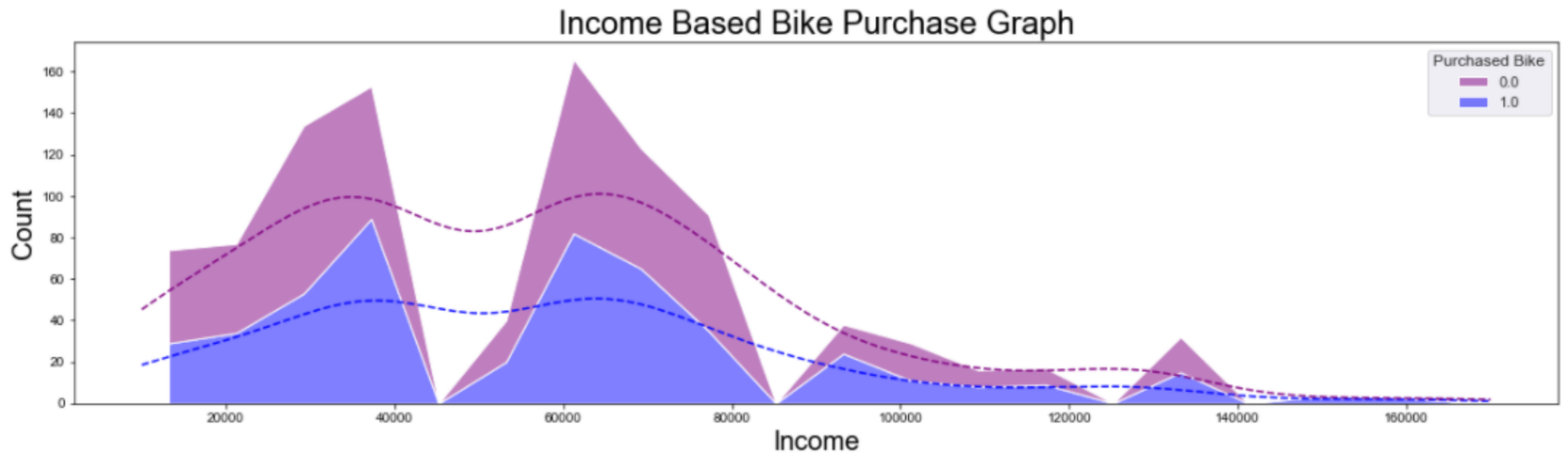
Education

The result shows that customers with higher education will buy the bike. Our hypothesis on this issue is that the use of bicycles actually makes a big difference in terms of protecting the environment, and the number of people with this awareness is directly proportional to the level of education. For this reason, it is seen that people with higher education levels buy more bicycles.

Income

The chart below shows that people with high incomes have a very low rate of buying bicycles.Our hypothesis regarding the reason for this is that high-income persons purchase primarily motor vehicles for transportation needs.

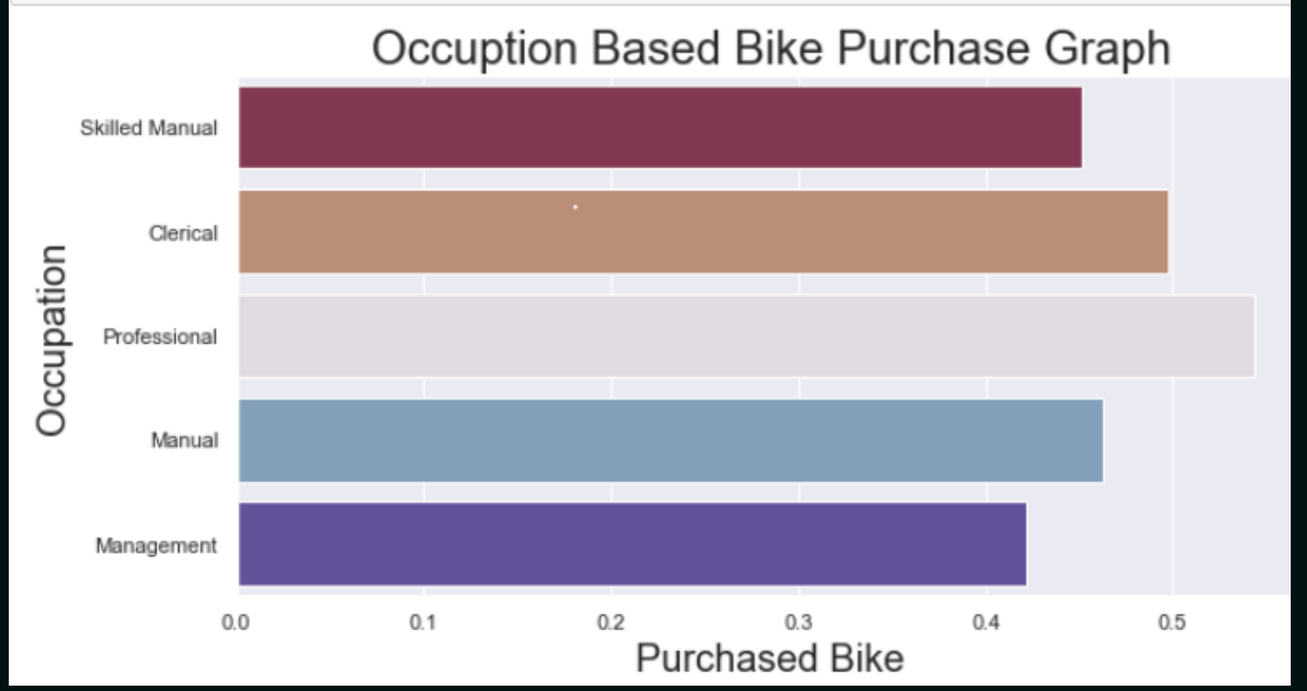
```
#create chart according to income
fig, ax=mplt.subplots(figsize=(20, 5))
sbrn.set_theme(style='darkgrid', palette=None)

#color and visual adjutments
visualB=sbrn.histplot(data=ds,x="Income", hue="Purchased Bike" ,shrink=0.8,kde=True,
          line_kws= {'linestyle': 'dashed'},multiple="stack", element="poly",
          palette=sbrn.blend_palette(['purple','blue'] ,2))

#physical property adjustment of the graph
visualB.set(title="Income Based Bike Purchase Graph")
visualB.xaxis.get_label().set_fontsize(20)
visualB.yaxis.get_label().set_fontsize(20)
visualB.title.set_fontsize(24)

#display
mplt.show()
```



Occuption

Based on the statistics, we discovered that professional employees are more likely than others to purchase a bicycle. However, despite this data, we see that the rate of bike purchase is not something that varies by occupation, the findings are relatively near to each other.

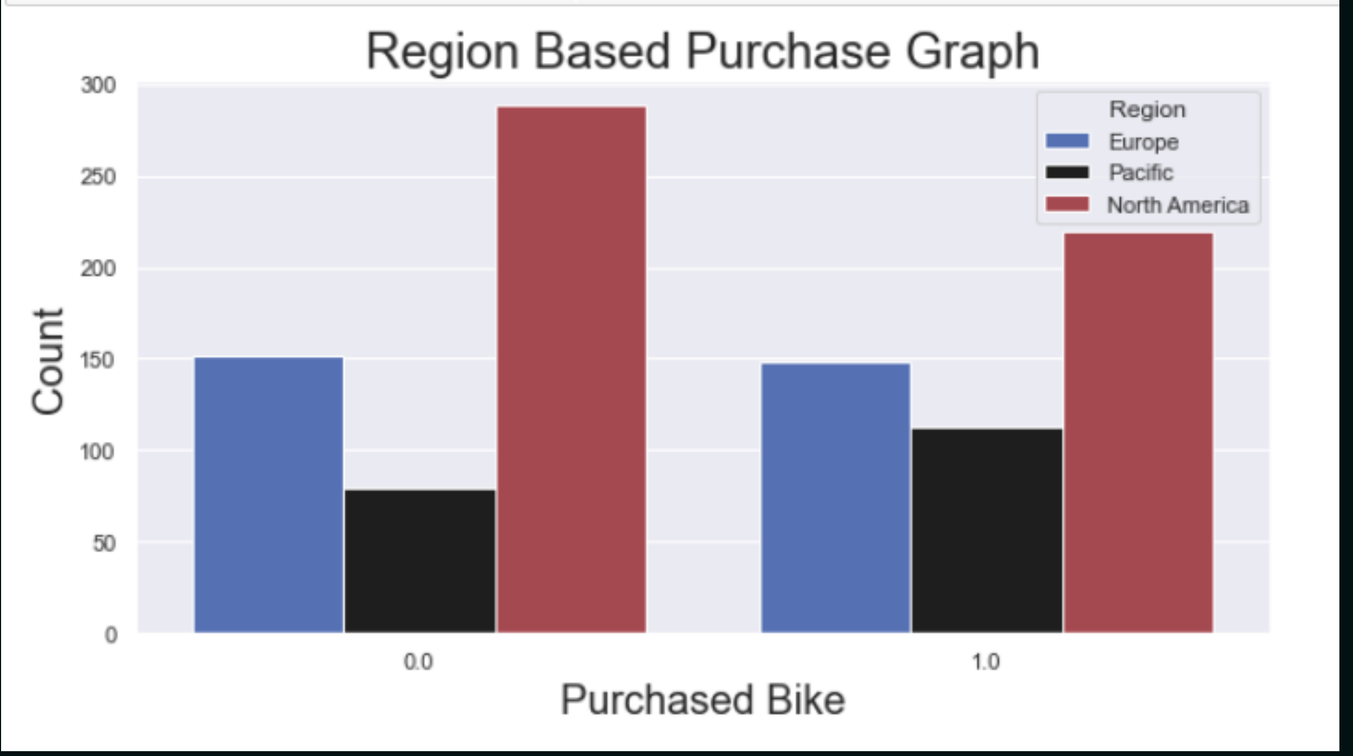
```
#create chart according to occuption
fig, cx=mplt.subplots(figsize=(10, 5))
visualC=sbrn.barplot(x = 'Purchased Bike', y = 'Occupation', data = ds, palette='twilight_shifted_r',ci=0)

#physical property adjustment of the graph
visualC.set(title="Occuption Based Bike Purchase Graph")
visualC.xaxis.get_label().set_fontsize(20)
visualC.yaxis.get_label().set_fontsize(20)
visualC.title.set_fontsize(24)
```



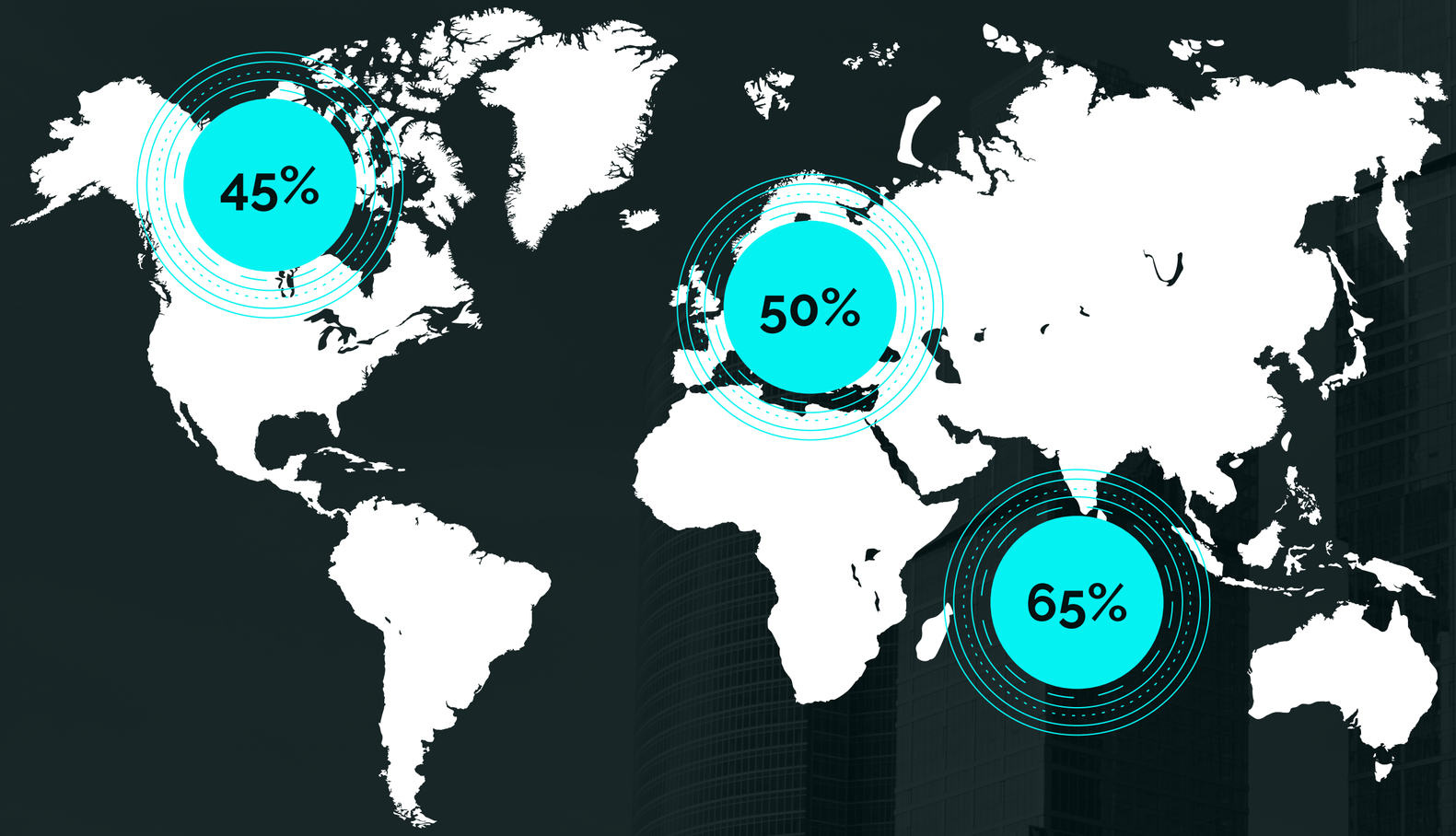
Region

The graph shows the number of people who buy bicycles and who don't by region. According to these data, (approx.) 50% of Europe, (approx.) 65% of Pacific and (approx.) 45% of North America have purchased bicycles.

```
#create chart according to region
fig, ax=mplt.subplots(figsize=(10, 5))
visualE=sbrn.countplot(x="Purchased Bike" ,data=ds, hue='Region', palette="icefire", saturation=0.7)

#physical property adjustment of the graph
visualE.set(title="Region Based Purchase Graph",ylabel='Count')
visualE.xaxis.get_label().set_fontsize(20)
visualE.yaxis.get_label().set_fontsize(20)
visualE.title.set_fontsize(24)

#display
mplt.show()
```



Although more bikes are sold in North America, the Pacific region has a higher purchase-to-population ratio.Our hypothesis in this regard is that customers in the pacific have higher purchasing power or the opportunity to ride a bike.

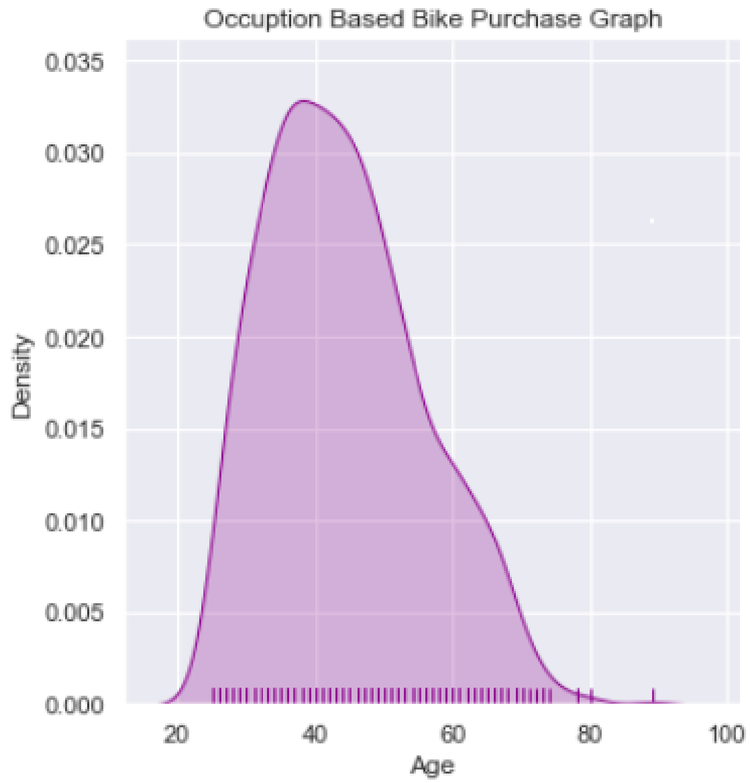


Age

According to the findings, younger clients are more likely to purchase a bike than older customers. Our comment on that is simple, health conditions. As the age increases, the chance of using a bicycle decreases due to natural reasons, and as a result, beyond a certain age, bicycle sales begin to decline.

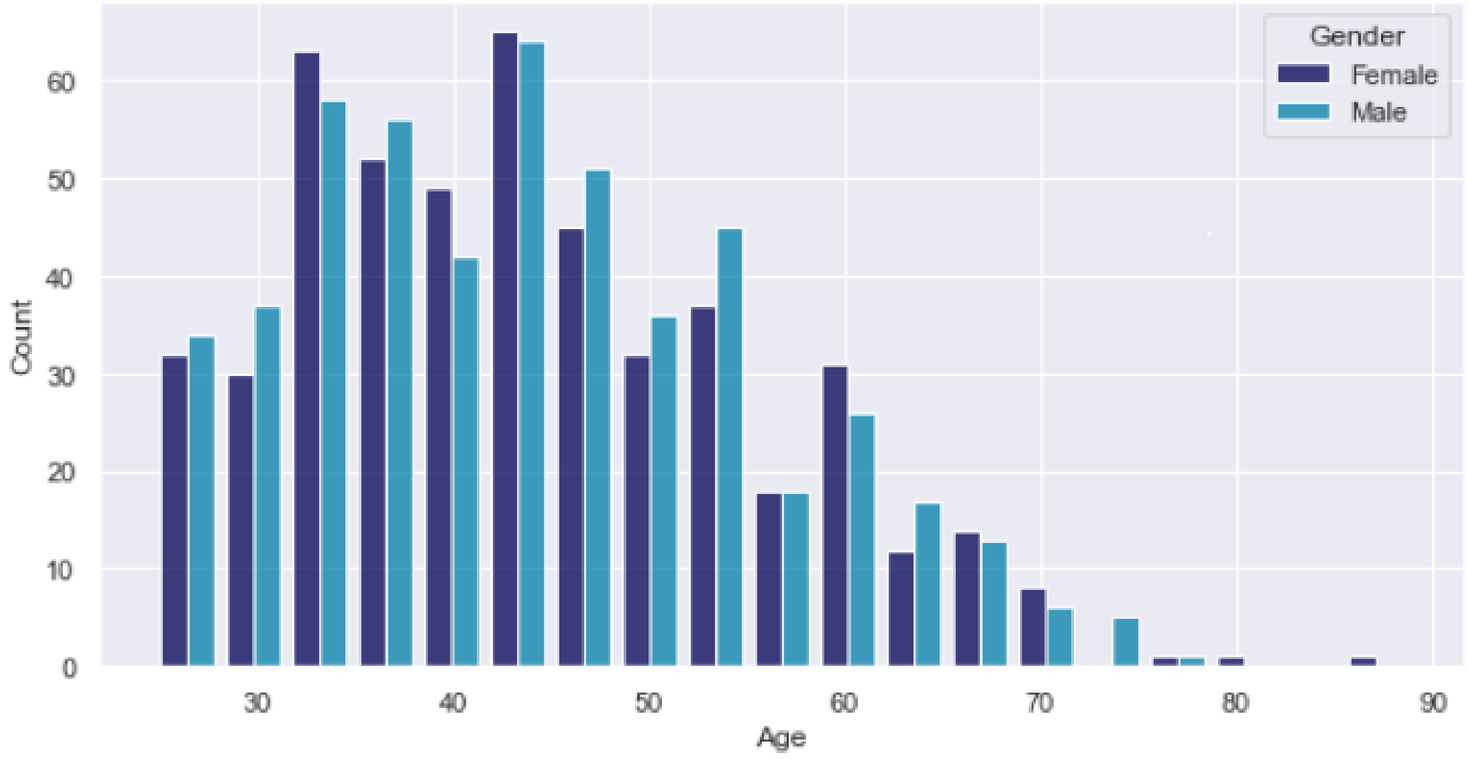
In [8]:
```
#create chart according to age
displot = ds['Age']
visualD=sbrn.displot(data=ds, x=displot, kind='kde', rug=True,fill=True, height=5, color="darkmagenta")

#title adjustment
visualD.set(title="Age Based Bike Purchase Graph")
```

Out[8]: <seaborn.axisgrid.FacetGrid at 0x242dec045e0>



Age vs. Gender

In [10]:
```
#create and resize chart that shows age vs gender
fig, dx=mplt.subplots(figsize=(10, 5))
visualF=sbrn.histplot(x = 'Age', hue='Gender', data = ds,shrink=0.8,multiple="dodge",palette="ocean")
```



Finally, in order to compare bicycle purchases by gender, we produced an age-based purchase graph. it's a quite detailed graph and shows that the bike purchase number of males are higher than the womens for the most of the time especialy after age 43.
ın this situation our hypothesis is that, in general men are more sportive and has better physical-health conditions rather than womens, particularly after middle ages.

PART VI

Which type of ML method have you used?

Classification

Which evaluation metric did you choose to evaluate your model?

Accuracy, but for improving, we also used confusion matrix,precision,recall ,f1 score by importing classification report.

For each of those algorithm, how did you choose appropriate parameters?

In Decision Tree, we used grid search for finding the best value for hyper-parameter.

Which algorithms did you use?

Naive Bayes and Decision Tree

PART VI

Did you select the features to be included in the model intelligently?

No we did not.

Did the the model performance change?

No.

Which feature selection methods did you use?

Supervised.

Which features are included in your final model?

- Visualed decision tree
- Accuracy score
- Confusion matrix

PART VII

RESULTS

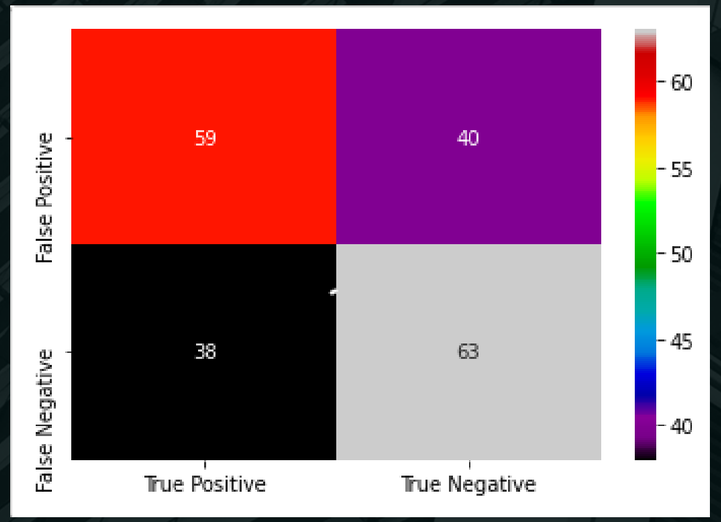
In this project, we see that data preprocessing especialy data cleaning plays an important role for the data scince projects. Secondly this implementation demonstrated that a wide variety of designs can be made for graphics with python and these visuals are critical for analysis and interpretation of the data. And finally, machine learning, Which we think is the most important part, takes up a lot of space in this project.

About our algorithms, Naive Bayes algorithm is easy and fast to predict class of test data set. When assumption of independence holds, a Naive Bayes classifier performs better compare to other models like logistic regression and you need less training data. Decision tree algorithm , usually mimic human thinking ability while making a decision, so it is easy to understand and also the logic behind the decision tree can be easily understood because it shows a tree-like structure.

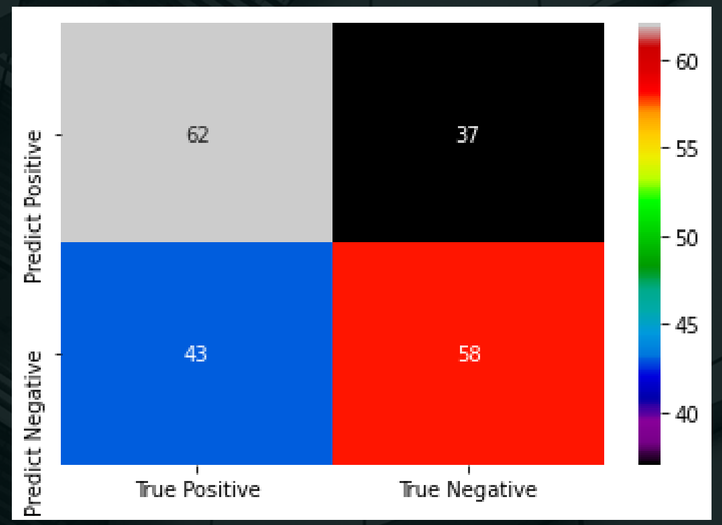
VISUALIZED RESULTS

	Accuracy score of model	Accuracy score of training set
Naive Bayes	0.6100	0.6225
Decision Tree	0.5650	0.6625

Naive Bayes



Decision Tree



Thanks For Watching