subjective-questions

December 17, 2023

Assignment-based Subjective Questions 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Ans-1. I have done an analysis of categorical variables from the data set on the dependent variable(count) using the boxplot and bar plot. Below are some points that can be inferred from the visualization;

a) fall has the highest median, which is expected as the weather conditions are most optimat to ride the bike followed by summer.
b) Median bike rents are increasing every year as the year 2019 has a higher median than 2018, it might be because Bike rentals are getting more popular and people are becoming more aware.
c) Most of the bikes were rented between May to October. This trend increased at the starting of the year till mid of the year, and then it started decreasing as we approach the year end.
d) People rent more on weekday compared to weekends, so the reason might be they prefer to spend time with family at home or use a personal vehicle.
e) Renting bikes seeems to be almost equal every day of the week.
f) Renting bikes seeems to be almost equal either on working days or non working days.
g) Clear weather is most optimal for bike renting, as temperature is optimal and humidity is less.

2. Why is it important to use drop_first=True during dummy variable creation?

Ans-2. A variable with n levels can be represented by n-1 dummy variable. So, if we remove the first column then also, we can represent the data. So, we use drop_first=True to reduce the extra column created during dummy variable creation. Hence it also reduces the correlations created among dummy variables.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Ans-3. By looking at the pairplot, it is quite clear that registered , casual , temp and atemp variables have the highest correlation with the target variable. But after dropping registered , casual, and atemp variables, the temp variable has the highest correlation with the target variable.

4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Ans-4. I have validated the assumptions of Linear Regression Model based on the below 5 assumptions -:

a) Normality of Error Terms: a)1) Error Terms should be normally distributed.

b) Multi-collinearity: b)1) There should be insignificant Multi-collinearity among variables.

c) Linear Relationship validation: c)1) Linearity should be visible among variables.

d) Homoscedasticity: d)1) There should be no visible pattern in residual values.

e) Independence of residuals: e)1) No Auto-correlation: " The Durbin-Watson value for final model is 2.0884. which means there is almost no auto-correlation.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Ans-5. Based on the coefficients of the final model, the top 3 features which explain the demand for the shared bikes are:

a) temp
b) light snow (weathersit (3))
c) yr (year)

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

Ans-1.Linear regression is a type of supervised machine learning algorithm that computes the linear relationship between a dependent variable and one or more independent features. When the number of the independent feature, is 1 then it is known as Univariate Linear regression, and in the case of more than one feature, it is known as multivariate linear regression.

Assumption for Linear Regression Model Linear regression is a powerful tool for understanding and predicting the behaviour of a variable, however, it needs to meet a few conditions in order to be accurate and dependable solutions.

Linearity: The independent and dependent variables have a linear relationship with one another. This implies that changes in the dependent variable follow those in the independent variable(s) in a linear fashion. This means that there should be a straight line that can be drawn through the data points. If the relationship is not linear, then linear regression will not be an accurate model.

Independence: The observations in the dataset are independent of each other. This means that the value of the dependent variable for one observation does not depend on the value of the dependent variable for another observation. If the observations are not independent, then linear regression will not be an accurate model.

Homoscedasticity: Across all levels of the independent variable(s), the variance of the errors is constant. This indicates that the amount of the independent variable(s) has no impact on the variance of the errors. If the variance of the residuals is not constant, then linear regression will not be an accurate model.

Normality: The residuals should be normally distributed. This means that the residuals should follow a bell-shaped curve. If the residuals are not normally distributed, then linear regression will not be an accurate model.

No multicollinearity: There is no high correlation between the independent variables. This indicates that there is little or no correlation between the independent variables.Multicollinearity occurs when two or more independent variables are highly correlated with each other, which can make it

difficult to determine the individual effect of each variable on the dependent variable. If there is multicollinearity, then linear regression will not be an accurate model.

Types of Linear Regression There are two main types of linear regression:

Simple Linear Regression: This is the simplest form of linear regression, and it involves only one independent variable and one dependent variable. The equation for simple linear regression is: y= *{0}*+ {1}X
where: Y is the dependent variable X is the independent variable 0 is the intercept 1 is the slope

Multiple Linear Regression: This involves more than one independent variable and one dependent variable. The equation for multiple linear regression is: y= *{0}*+ {1}X+ *{2}X*+…… {n}X
where: Y is the dependent variable X1, X2, …, Xp are the independent variables 0 is the intercept 1, 2, …, n are the slopes

[ ]:
```
2. Explain the Anscombe's quartet in detail. (3 marks)

Ans-2.Anscombe's quartet comprises a set of four dataset, having identical␣
 ↪descriptive statistical properties in terms of means, variance, R-Squared,␣
 ↪correlations, and linear regression lines but having different␣
 ↪representations when we scatter plot on graph. The datasets were created by␣
 ↪the statistician Francis Anscombe in 1973 to demonstrate the importance of␣
 ↪visualizing data and to show that summary statistics alone can be misleading.

The four datasets that make up Anscombe's quartet each include 11 x-y pairs of␣
 ↪data. When plotted, each dataset seems to have a unique connection between x␣
 ↪and y, with unique variability patterns and distinctive correlation␣
 ↪strengths. Despite these variations, each dataset has the same summary␣
 ↪statistics, such as the same x and y mean and variance, x and y correlation␣
 ↪coefficient, and linear regression line.

Anscombe's quartet is used to illustrate the importance  of exploratory data␣
 ↪analysis and the drawbacks of depending only on summary statistics.  It also␣
 ↪emphasizes the importance of using data visualization to spot trends,␣
 ↪outliers, and other crucial details that might not be obvious from summary␣
 ↪statistics alone.

The four datasets of Anscombe's quartet.
```

[1]:
```
# Import the necessary libraries
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
```

[3]:
```
df = pd.read_csv('anscombe.csv')
print(df)
```

```
    x1  x2  x3  x4     y1    y2     y3     y4
0   10  10  10   8   8.04  9.14   7.46   6.58
```

1	8	8	8	8	6.95	8.14	6.77	5.76
2	13	13	13	8	7.58	8.74	12.74	7.71
3	9	9	9	8	8.81	8.77	7.11	8.84
4	11	11	11	8	8.33	9.26	7.81	8.47
5	14	14	14	8	9.96	8.10	8.84	7.04
6	6	6	6	8	7.24	6.13	6.08	5.25
7	4	4	4	19	4.26	3.10	5.39	12.50
8	12	12	12	8	10.84	9.13	8.15	5.56
9	7	7	7	8	4.82	7.26	6.42	7.91
10	5	5	5	8	5.68	4.74	5.73	6.89

Find the descriptive statistical properties for the all four dataset Find mean for x and y for all four datasets. Find standard deviations for x and y for all four datasets. Find correlations with their corresponding pair of each datasets. Find slope and intercept for each datasets. Find R-square for each datasets. To find R-square first find residual sum of square error and Total sum of square error Create a statistical summary by using all these data and print it

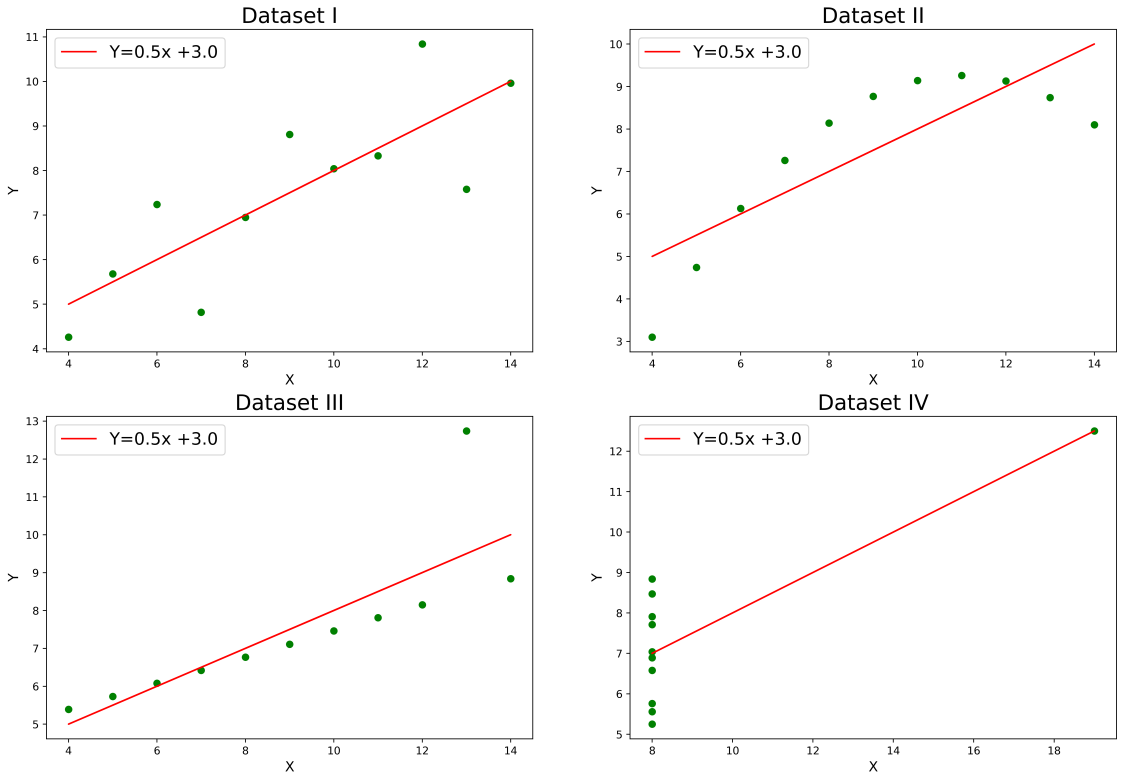
[4]:
```
# mean values (x-bar)
x1_mean = df['x1'].mean()
x2_mean = df['x2'].mean()
x3_mean = df['x3'].mean()
x4_mean = df['x4'].mean()

# y-bar
y1_mean = df['y1'].mean()
y2_mean = df['y2'].mean()
y3_mean = df['y3'].mean()
y4_mean = df['y4'].mean()


# Standard deviation values (x-bar)
x1_std = df['x1'].std()
x2_std = df['x2'].std()
x3_std = df['x3'].std()
x4_std = df['x4'].std()

# Standard deviation values (y-bar)
y1_std = df['y1'].std()
y2_std = df['y2'].std()
y3_std = df['y3'].std()
y4_std = df['y4'].std()

# Correlation
correlation_x1y1 = np.corrcoef(df['x1'],df['y1'])[0,1]
correlation_x2y2 = np.corrcoef(df['x2'],df['y2'])[0,1]
correlation_x3y3 = np.corrcoef(df['x3'],df['y3'])[0,1]
correlation_x4y4 = np.corrcoef(df['x4'],df['y4'])[0,1]
```

```
# Linear Regression slope and intercept
m1,c1 = np.polyfit(df['x1'], df['y1'], 1)
m2,c2 = np.polyfit(df['x2'], df['y2'], 1)
m3,c3 = np.polyfit(df['x3'], df['y3'], 1)
m4,c4 = np.polyfit(df['x4'], df['y4'], 1)

# Residual sum of squares error
RSSY_1 = ((df['y1'] - (m1*df['x1']+c1))**2).sum()
RSSY_2 = ((df['y2'] - (m2*df['x2']+c2))**2).sum()
RSSY_3 = ((df['y3'] - (m3*df['x3']+c3))**2).sum()
RSSY_4 = ((df['y4'] - (m4*df['x4']+c4))**2).sum()

# Total sum of squares
TSS_1 = ((df['y1'] - y1_mean)**2).sum()
TSS_2 = ((df['y2'] - y2_mean)**2).sum()
TSS_3 = ((df['y3'] - y3_mean)**2).sum()
TSS_4 = ((df['y4'] - y4_mean)**2).sum()

# R squared (coefficient of determination)
R2_1  = 1 - (RSSY_1 / TSS_1)
R2_2  = 1 - (RSSY_2 / TSS_2)
R2_3  = 1 - (RSSY_3 / TSS_3)
R2_4  = 1 - (RSSY_4 / TSS_4)

# Create a pandas dataframe to represent the summary statistics
summary_stats = pd.DataFrame({'Mean_x': [x1_mean, x2_mean, x3_mean, x4_mean],
                              'Variance_x': [x1_std**2, x2_std**2, x3_std**2,␣
 ↪x4_std**2],
                              'Mean_y': [y1_mean, y2_mean, y3_mean, y4_mean],
                              'Variance_y': [y1_std**2, y2_std**2, y3_std**2,␣
 ↪y4_std**2],
                              'Correlation': [correlation_x1y1,␣
 ↪correlation_x2y2, correlation_x3y3, correlation_x4y4],
                              'Linear Regression slope': [m1, m2, m3, m4],
                              'Linear Regression intercept': [c1, c2, c3, c4]},
index = ['I', 'II', 'III', 'IV'])
print(summary_stats.T)
```

```
                                I         II        III         IV
Mean_x                   9.000000   9.000000   9.000000   9.000000
Variance_x              11.000000  11.000000  11.000000  11.000000
Mean_y                   7.500909   7.500909   7.500000   7.500909
Variance_y               4.127269   4.127629   4.122620   4.123249
Correlation              0.816421   0.816237   0.816287   0.816521
Linear Regression slope  0.500091   0.500000   0.499727   0.499909
Linear Regression intercept  3.000091   3.000909   3.002455   3.001727
```

[5]:
```
# Plot the scatter plot and linear regression line for each datasets
```

[6]:
```
# plot all four plots
fig, axs = plt.subplots(2, 2,  figsize=(18,12), dpi=500)

axs[0, 0].set_title('Dataset I', fontsize=20)
axs[0, 0].set_xlabel('X', fontsize=13)
axs[0, 0].set_ylabel('Y', fontsize=13)
axs[0, 0].plot(df['x1'], df['y1'], 'go')
axs[0, 0].plot(df['x1'], m1*df['x1']+c1,'r',label='Y='+str(round(m1,2))+'x␣
 ↪+'+str(round(c1,2)))
axs[0, 0].legend(loc='best',fontsize=16)

axs[0, 1].set_title('Dataset II',fontsize=20)
axs[0, 1].set_xlabel('X', fontsize=13)
axs[0, 1].set_ylabel('Y', fontsize=13)
axs[0, 1].plot(df['x2'], df['y2'], 'go')
axs[0, 1].plot(df['x2'], m2*df['x2']+c2,'r',label='Y='+str(round(m2,2))+'x␣
 ↪+'+str(round(c2,2)))
axs[0, 1].legend(loc='best',fontsize=16)

axs[1, 0].set_title('Dataset III',fontsize=20)
axs[1, 0].set_xlabel('X', fontsize=13)
axs[1, 0].set_ylabel('Y', fontsize=13)
axs[1, 0].plot(df['x3'], df['y3'], 'go')
axs[1, 0].plot(df['x3'], m1*df['x3']+c1,'r',label='Y='+str(round(m3,2))+'x␣
 ↪+'+str(round(c3,2)))
axs[1, 0].legend(loc='best',fontsize=16)

axs[1, 1].set_title('Dataset IV',fontsize=20)
axs[1, 1].set_xlabel('X', fontsize=13)
axs[1, 1].set_ylabel('Y', fontsize=13)
axs[1, 1].plot(df['x4'], df['y4'], 'go')
axs[1, 1].plot(df['x4'], m4*df['x4']+c4,'r',label='Y='+str(round(m4,2))+'x␣
 ↪+'+str(round(c4,2)))
axs[1, 1].legend(loc='best',fontsize=16)

plt.show()
```



[7]:
```
#### Note: It is mentioned in the definition that Anscombe's quartet comprises␣
 ↪four datasets that have nearly identical simple statistical properties, yet␣
 ↪appear very different when graphed.

### Explanation of this output:
# In the first one(top left) if you look at the scatter plot you will see that␣
 ↪there seems to be a linear relationship between x and y.
# In the second one(top right) if you look at this figure you can conclude that␣
 ↪there is a non-linear relationship between x and y.
# In the third one(bottom left) you can say when there is a perfect linear␣
 ↪relationship for all the data points except one which seems to be an outlier␣
 ↪which is indicated be far away from that line.
# Finally, the fourth one(bottom right) shows an example when one high-leverage␣
 ↪point is enough to produce a high correlation coefficient.
```

3. What is Pearson's R?

Ans-3.
In Statistics, the Pearson's Correlation Coefficient is also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or bivariate correlation. It is a statistic that measures the linear correlation between two variables. Like all correlations, it also has a numerical value that lies between -1.0 and +1.0. Whenever we discuss correlation in statistics, it is generally Pearson's correlation coefficient. However, it cannot capture nonlinear relationships between two

variables and cannot differentiate between dependent and independent variables. Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations. The form of the definition involves a "product moment", that is, the mean (the first moment about the origin) of the product of the mean-adjusted random variables; hence the modifier product-moment in the name. Pearson's Correlation Coefficient is named after Karl Pearson. He formulated the correlation coefficient from a related idea by Francis Galton in the 1880s.

Pearson's Correlation Coefficient Formula

The Pearson's correlation coefficient formula is the most commonly used and the most popular formula to get the correlation coefficient. It is denoted with the capital "R". The formula for Pearson's correlation coefficient is shown below,

$R= n(\sum xy) - (\sum x)(\sum y) / \sqrt{[n \sum x^2-(\sum x)^2][n \sum y^2-(\sum y)^2}$

The full name for Pearson's correlation coefficient formula is Pearson's Product Moment correlation (PPMC). It helps in displaying the Linear relationship between the two sets of the data.

The Pearson's correlation helps in measuring the strength(it's given by coefficient r-value between -1 and +1) and the existence (given by p-value)of a linear relationship between the two variables and if the outcome is significant we conclude that the correlation exists.

Cohen (1988) says that an absolute value of r of 0.5 is classified as large, an absolute value of 0.3 is classified as medium and an absolute value of 0.1 is classified as small.

The interpretation of the Pearson's correlation coefficient is as follows:-

A correlation coefficient of 1 means there is a positive increase of a fixed proportion of others, for every positive increase in one variable. Like, the size of the shoe goes up in perfect correlation with foot length. If the correlation coefficient is 0, it indicates that there is no relationship between the variables. A correlation coefficient of -1 means there is a negative decrease of a fixed proportion, for every positive increase in one variable. Like, the amount of water in a tank will decrease in a perfect correlation with the flow of a water tap.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Ans-4. It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude. It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

Normalization/Min-Max Scaling: It brings all of the data in the range of 0 and 1. sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.

Standardization Scaling: Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean ( ) zero and standard deviation one ( ).

sklearn.preprocessing.scale helps to implement standardization in python. One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Ans-5. Variance Inflation Factor or VIF, is a measure of the amount of multi-collinearity in a set of multiple regression variables.

The formula for the variance inflation factor is given as : VIF = 1/(1-R(square)) Now, when calculating the VIF for one independent variable using all the other independent variables. if the r-square value comes out to be 1, the VIF value will become infinite. This is quite possible when one of the independent variables is strongly correlated with many of the other independent variables.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Ans-6. The quantile-quantile plot is a graphical method for determining whether two samples of data came from the same population or not. A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. By a quantile, we mean the fraction (or percent) of points below the given value. For the reference purpose, a 45% line is also plotted, if the samples are from the same population then the points are along this line.

Normal Distribution: The normal distribution (aka Gaussian Distribution/ Bell curve) is a continuous probability distribution representing distribution obtained from the randomly generated real values.

Normal Distribution with Area Under CUrve

Usage: The Quantile-Quantile plot is used for the following purpose:

Determine whether two samples are from the same population. Whether two samples have the same tail Whether two samples have the same distribution shape. Whether two samples have common location behavior.

How to Draw Q-Q plot Collect the data for plotting the quantile-quantile plot. Sort the data in ascending or descending order. Draw a normal distribution curve. Find the z-value (cut-off point) for each segment. Plot the dataset values against the normalizing cut-off points.

Advantages of Q-Q plot Since Q-Q plot is like probability plot. So, while comparing two datasets the sample size need not to be equal. Since we need to normalize the dataset, so we don't need to care about the dimensions of values.