Assignment-based Subjective Questions

Q1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

Answer:

1. Season has effect on bike sales as higher sale in Fall & lower in spring.
2. Months also effect sales as, May to Sep sales in peak, lower in Jan, Feb & Dec.
3. Days has no effect on sales, looks almost similar.
4. Weekdays has no effect on sales, looks almost similar
5. Holiday effects sales, on holidays sale is low compare to non-holiday.
6. Working-day day have no significance difference on sale.
7. Weathersit, when weather is clear higher is sale & low in light snow.

Q2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)

Answer:

To overcome the curse of dimensionality, we set drop_first=T, its like for three segment A, B & C.

If A = 0, B =0 then its automatically understandable that C = 1.

Q3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Answer:

By looking at pair plots, we can see scatter plot showing distribution on points, we distributions of points, and we are seeing pattern like 45 Degree upward or downward (or nearby) of points with target variable, conclude that variable is highly correlated (Positively or Negatively) with target. We can also check correlation by using corr() function. -1 means negative correlation, 0 no correlation, 1 positive correlation.

Q4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 Marks)

Answer:

To check the assumption of linear regression.

1. We check Residuals are normally distributed, by plotting histogram on residuals.
2. No sign of heteroscedasticity, by plotting scatter plot on residuals.
3. Predicted outcome normally distributed, by plotting histogram on predicted values.

Q5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 Marks)

Answer:

Top 3 contributor are:

1. Year
2. Light Snow
3. Spring

General Subjective Questions

Q1. Explain the linear regression algorithm in detail. (4 Marks)

Answer:

Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression.

Equation of multiple linear regression is like:

Where bo is the intercept & $b_1,b_2,b_3,b_4\ldots,b_n$ are coefficients or slopes of the independent variables $x_1,x_2,x_3,x_4\ldots,x_n$ and y is the dependent variable.

$$y= b_o + b_1x_1 + b_2x_2 + b_3x_3 \ldots. + b_nx_n$$

We need to to best fit line with OLS, having lowest sum of residuals.

Metrics for model evaluation

R-Squared value

This value ranges from 0 to 1. Value '1' indicates predictor perfectly accounts for all the variation in Y. Value '0' indicates that predictor 'x' accounts for no variation in 'y'.

1. Regression sum of squares (SSR)

This gives information about how far estimated regression line is from the horizontal 'no relationship' line (average of actual output).

$$\text{Error} = \sum_{i=1}^{n} (\text{Predicted\_output} - \text{average\_of\_actual\_output})\text{^}2$$

Figure 9: Regression Error Formula

2. Sum of Squared error (SSE)

How much the target value varies around the regression line (predicted value).

$$\text{Error} = \sum_{i=1}^{n} (Actual\_output - predicted\_output)\text{^}2$$

Figure 10: Sum of Square Formula

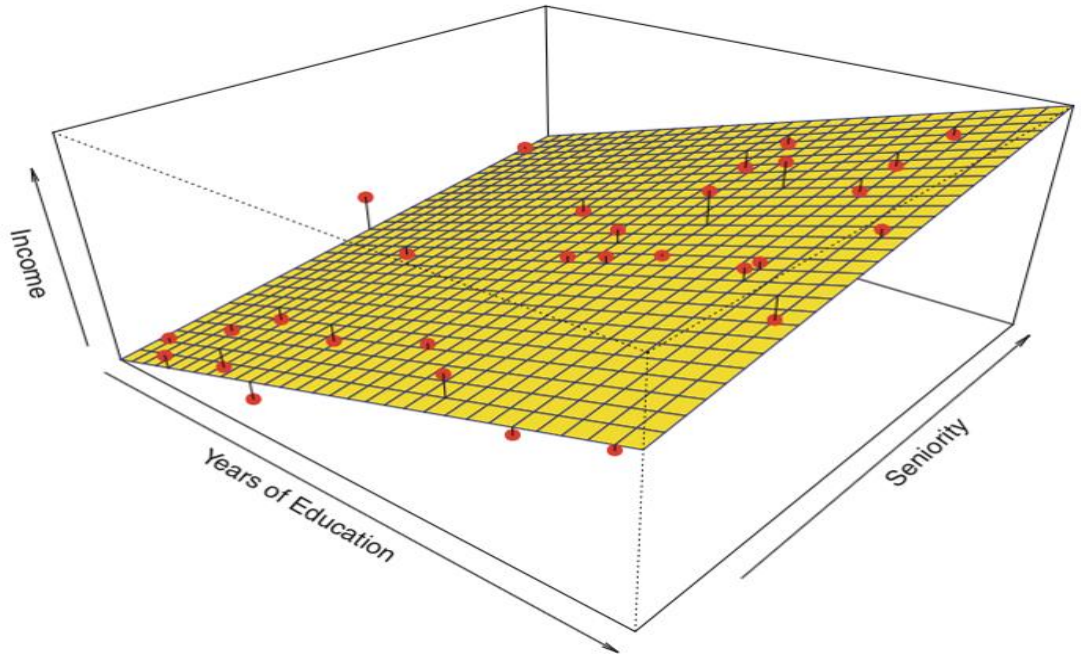
3. Total sum of squares (SSTO)

This tells how much the data point move around the mean.

$$\text{Error} = \sum_{i=1}^{n} (\text{Actual\_output} - \text{average\_of\_actual\_output})\text{^}2$$

R^2 = 1 – (SSE/SSTO)

We need to fit the plan in case of multiple linear regression, having lowest residual error.



Q2. Explain the Anscombe's quartet in detail. (3 Marks)

Answer:

Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers on statistical properties.

Simple understanding:

Once Francis John "Frank" Anscombe who was a statistician of great repute found 4 sets of 11 data-points in his dream and requested the council as his last wish to plot those points. Those 4 sets of 11 data-points are given below.

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

After that, the council analyzed them using only descriptive statistics and found the mean, standard deviation, and correlation between x and y.

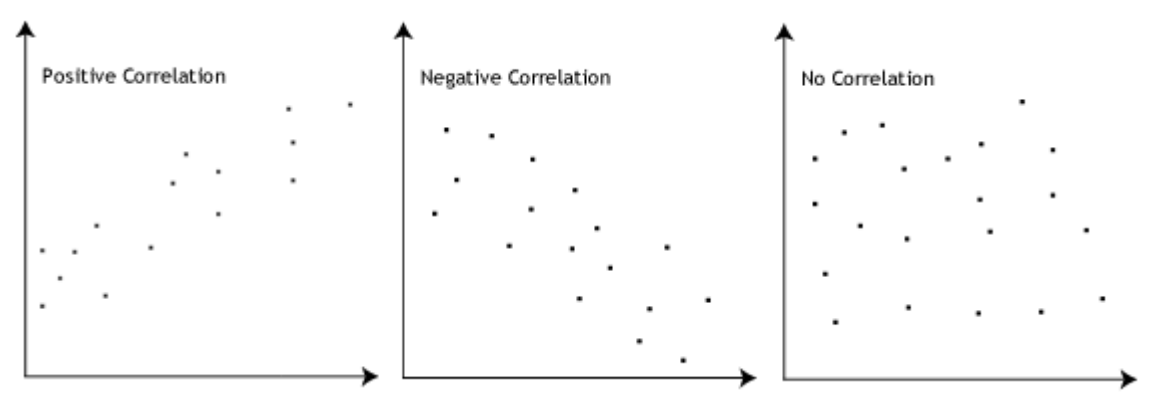
Q3. What is Pearson's R? (3 Marks)

Answer:

In statistics, the Pearson correlation coefficient (PCC), also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or the bivariate correlation, is a measure of linear correlation between two sets of data. It is the covariance of two variables, divided by the product of their standard deviations; thus it is essentially a normalised measurement of the covariance, such that the result always has a value between −1 and 1.

The Pearson's correlation coefficient varies between -1 and +1 where:

- r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)
- r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)
- r = 0 means there is no linear association
- r > 0 < 5 means there is a weak association
- r > 5 < 8 means there is a moderate association
- r > 8 means there is a strong association



Pearson r Formula

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Here,

- =correlation coefficient
- =values of the x-variable in a sample
- =mean of the values of the x-variable

- =values of the y-variable in a sample
- =mean of the values of the y-variable

Q4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 Marks)

Answer:

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.

Q5. 5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 Marks)

Answer:

If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

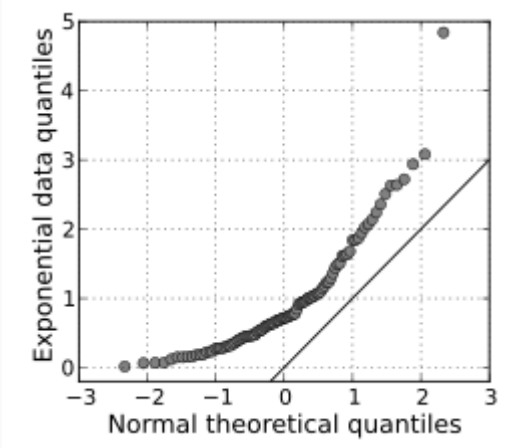
An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

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

Answer:

Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q Q plots is to find out if two sets of data come from the same distribution. A 45 degree angle is plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.

A Q Q plot showing the 45 degree reference line:



If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.

A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.