

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?

Answer:

Below are listed inferences I made from my analysis,

- Fall has the highest average rentals, which is expected as weather conditions are most optimal to ride a bike followed by summer.
- The year had a median value increase of up to 6000 in rentals compared to 2018. Hence Bikes rented are more in the year 2019 as compared to 2018. It might be due to the fact that bike rentals are getting into a trend
- Bikes rented are more in the month of September. It seems like the trend is explained by seasonal rentals too.
- Holidays show a lower rental count compared to working days.
- Weekends and Holidays have equal median value, although the spread is bigger for non-working as people might have a plan & do not want to rent bikes.
- There is no significant difference between weekdays, except that Thursdays and Sundays have a higher variation in rentals than others.
- Clear weather shows higher rentals compare to other weather. Because of optimal humidity, and less temperature.

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

Answer:

If we don't use `drop_first=True` you will get a redundant feature. As it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

Example: We have three variables: Furnished, Semi-furnished, and un-furnished. We can only take 2 variables as furnished will be 1-0, semi-furnished will be 0-1, so we don't need unfurnished as we know 0-0 will indicate un-furnished. So we can remove it.

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

Answer:

temp and atemp both have the same correlation(0.63) with the target variable which is the highest among all numerical variables.

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

Answer:

With below-listed points we can achieve the assumptions of Liner Regression

- The simple way to determine if this assumption is met or not is by creating a scatter plot x vs y. If the data points fall on a straight line in the graph, there is a linear relationship between the dependent and the independent variables, and the assumption holds otherwise not.

- There should be no correlation between the residual (error) terms. Absence of this phenomenon is known as Autocorrelation.
 - The independent variables should not be correlated. Absence of this phenomenon is known as multicollinearity.
 - The error terms must have constant variance. This phenomenon is known as homoskedasticity. The presence of non-constant variance is referred to heteroskedasticity.
 - The error terms must be normally distributed.
5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of shared bikes?**

Answer:

The top 3 features contributing significantly to explaining the demand for shared bikes are

1. **Temp:** With positive coefficients of 5195.6518
2. **Yr:** With positive coefficients of 1981.2753
3. **Winter:** With positive coefficients of 1179.7188

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**

Answer:

Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task, where we train a model to predict the behavior of your data based on some variables.

Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (x) variables, hence called as linear regression. Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable.

Linear regression makes predictions for continuous/real or numeric variables such as **sales, salary, age, product price**, etc.

Mathematically, we can write a linear regression equation as:

$$y = a + bx$$

Here, x and y are two variables on the regression line.

b = Slope of the line

a = y-intercept of the line

x = Independent variable from dataset

y = Dependent variable from dataset

2. Explain the Anscombe's quartet in detail.

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.

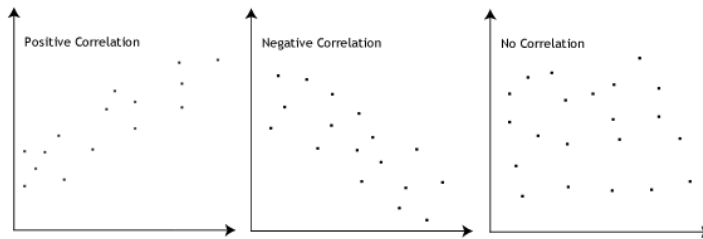
3. What is Pearson's R?

Answer:

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

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

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

Answer:

Scaling 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.

It is performed because 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.

$$\text{MinMax Scaling: } x = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Standardization Scaling:

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

$$\text{Standardisation: } x = \frac{x - \text{mean}(x)}{\text{sd}(x)}$$

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

Answer:

If all the independent variables are orthogonal to each other, then $VIF = 1.0$. If there is perfect correlation, then $VIF = \text{infinity}$. A large value of VIF indicates that there is a correlation between the variables. In the case of perfect correlation, we get $R^2 = 1$, which leads to $1/(1-R^2)$ 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).

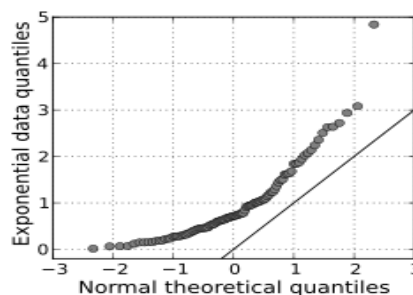
6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

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.