Assignment questions

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

Answer:

We can infer the below-

- Demand for bikes is less in spring as compared to other seasons
- 2019 shows increased demand
- The demand for rental bikes show increase from April till October and then gradually decreases November onwards
- Non holiday has higher rental bike counts
- Weekday show no significance in bike rentals, the median remains same across the days of week
- Similarly for working day no difference
- Demands are higher when weather is clear/partially cloudy

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

Answer:

drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it 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?

Answer:

We can see temp, atemp, yr and fall being top 4 highly correlated variables with cnt

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

Answer:

The assumptions of linear regression were validated by analysing the below:

- The residuals are normally distributed - Using distplot
- The residuals have mean=0 – Calculating mean of residuals
- The variance is constant i.e. no heteroscedasticity- Plotting residuals with predicted values
- No multicollinearity (All VIFs<5)- Checking VIF after final model

- Linear relationship between independent and dependent variable- Using pairplot

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

Answer:

Based on the final model the top 3 variables are:

1. Year 2019(yr)
2. Aug month
3. Light Snow (weathersit)

General Subjective questions

1. Explain the linear regression algorithm in detail

    Answer:

    Linear regression is one of the very basic forms of machine learning where we train a model to predict the behaviour of your data based on some variables. In the case of linear regression as you can see the name suggests linear that means the two variables which are on the x-axis and y-axis should be linearly correlated.

    Mathematically, we can write a linear regression equation as:
    y = a + bx

    Where a and b given by the formulas:

    b(slope) =

    $$\frac{n \sum xy - (\sum x)(\sum y)}{n \sum x^2 - (\sum x)^2}$$

    a(intercept)=

    $$\frac{n \sum y - b(\sum x)}{n}$$

    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	8,04	10	9,14	10	7,46	8	6,58
	8	6,95	8	8,14	8	6,77	8	5,76
	13	7,58	13	8,74	13	12,74	8	7,71
	9	8,81	9	8,77	9	7,11	8	8,84
	11	8,33	11	9,26	11	7,81	8	8,47
	14	9,96	14	8,1	14	8,84	8	7,04
	6	7,24	6	6,13	6	6,08	8	5,25
	4	4,26	4	3,1	4	5,39	19	12,5
	12	10,84	12	9,13	12	8,15	8	5,56
	7	4,82	7	7,26	7	6,42	8	7,91
	5	5,68	5	4,74	5	5,73	8	6,89
SUM	99,00	82,51	99,00	82,51	99,00	82,50	99,00	82,51
AVG	9,00	7,50	9,00	7,50	9,00	7,50	9,00	7,50
STDEV	3,32	2,03	3,32	2,03	3,32	2,03	3,32	2,03

The summary statistics show that the means and the variances were identical for x and y across the groups :
- Mean of x is 9 and mean of y is 7.50 for each dataset.
- Similarly, the variance of x is 11 and variance of y is 4.13 for each dataset
- The correlation coefficient (how strong a relationship is between two variables) between x and y is 0.816 for each dataset

When we plot these four datasets on an x/y coordinate plane, we can observe that they show the same regression lines as well but each dataset is telling a different story :
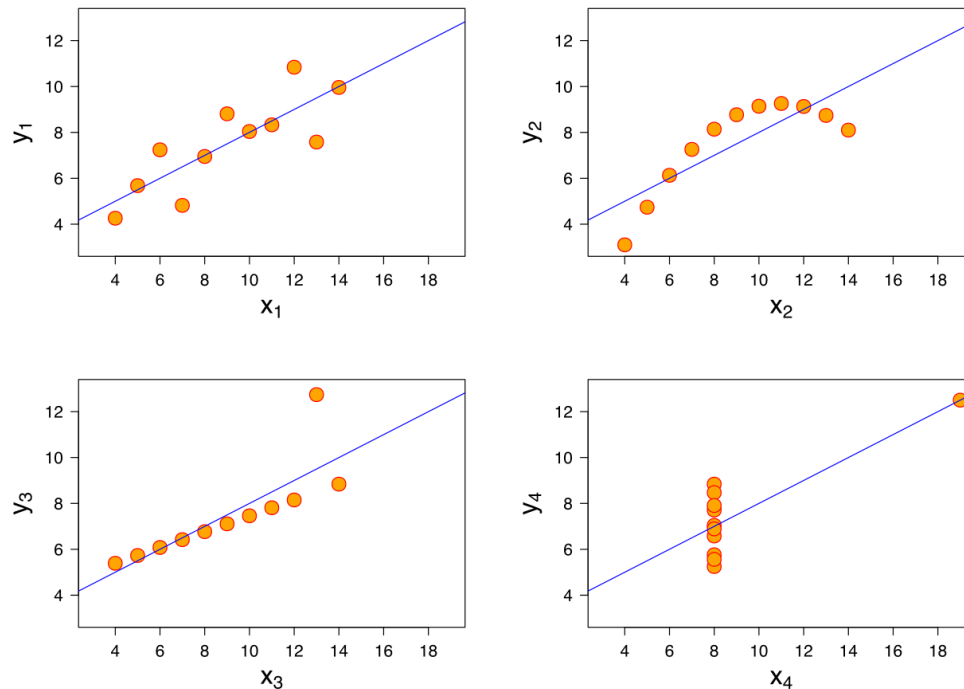
- Dataset I appears to have clean and well-fitting linear models.

- Dataset II is not distributed normally.

- In Dataset III the distribution is linear, but the calculated regression is thrown off by an outlier.

- Dataset IV shows that one outlier is enough to produce a high correlation coefficient.

This quartet emphasizes the importance of visualization in Data Analysis. Looking at the data reveals a lot of the structure and a clear picture of the dataset.

3. What is Pearson's R?

Answer:

- 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.
- Pearson's r is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation

coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.

- 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.

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

   Scaling:
   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.
   Why to perform scaling:
   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.
   Difference between Normalized and Standard scaling:
   Normalization typically means rescales the values into a range of [0,1]. Standardization typically means rescales data to have a mean of 0 and a standard deviation of 1 (unit variance).

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

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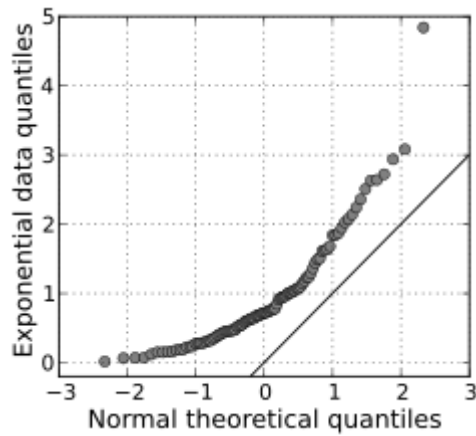
If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x. Q–Q plots can also be used as a graphical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