Assignment-based Subjective Questions

Question: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 below observation:

- Over 5K booking is happening on the Season3, Season2 and Season 4 whereas we observe that less than 3.5k booking is happening on season 1
- Over 4k boking happening on the range between 4 to 10.
- Over 4k booking is only happening in weathersit 1.
- most of the Bike booking happening when thers is a working day
- Weekday is independent, All the days have marginly same count

Question:2 Why is it important to use drop_first=True during dummy variable creation?
Answer:
drop_first=True,is important because importance or value of that left over variable can be found by remaining variables .So to avoid redundancy we are dropping a column. for ex: I have 3 friends Neha, Niti and Nitya and if i am introducing Neha and Niti to my brother, then automatically my brother knows who is Nitya. So once we create dummies for n-1 categories it's definite what the nth category is

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

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

- Residuals are normally distributed
- There is No Multicollinearity between the predictor variables
- There is a linear relationship between temp, atemp and cnt

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

- temp
- weathersit_3 (Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds)
- yr

General Subjective Questions

Question: Explain the linear regression algorithm in detail.
Answer:
Linear regression is supervised technique, which is used for predictive analysis. It shows the linear relationship between a dependent(y) and an independent (x) variable using a straight line, which means it explains how the value of the dependent variable is changing according to the value of the independent variable.

Mathematically, we can represent a linear regression as:
$y = bx + a$
Here,
x = independent variable (Predictor Variable)
y = dependent variable (Target Variable)
b = slope of the line
a = y-intercept of the line
Below are formulas for calculating a and b are:

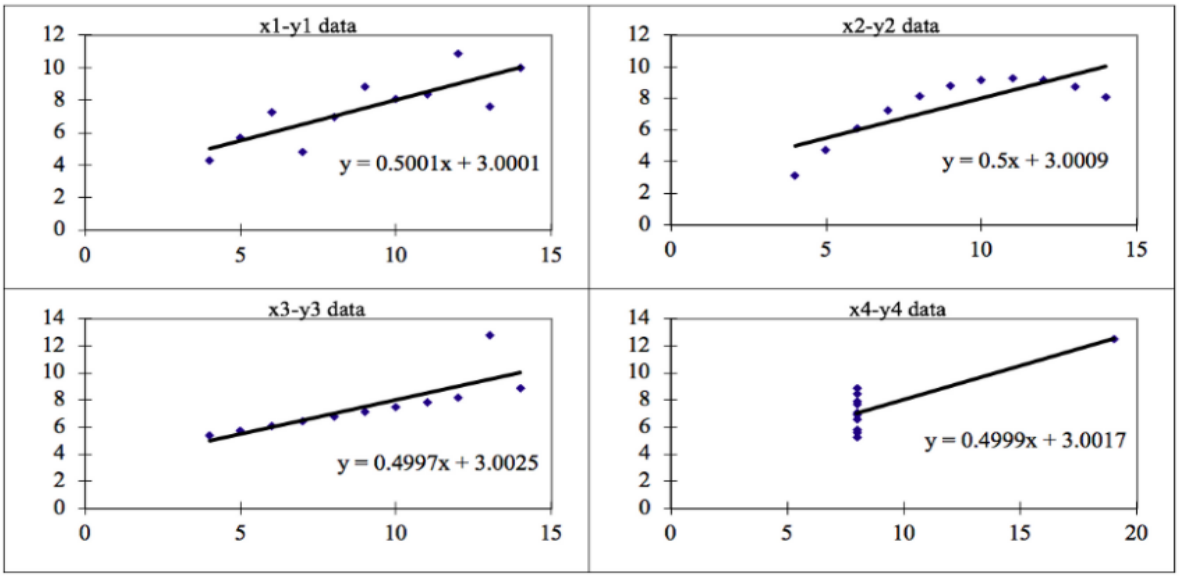
$$b(slobe) = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$$

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

Question: Explain the Anscombe's quartet in detail.
Answer:
Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics (mean, variance, etc), but there are some peculiarities in the dataset that fools the regression model if built. They also have very different distributions and appear differently when graphed.
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 the importance of plotting the graphs before analyzing and model building, and the effect of other observations on statistical properties
For Example:

Anscombe's Data										
Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
8	4	4.26		4	3.1		4	5.39		19	12.5
9	12	10.84		12	9.13		12	8.15		8	5.56
10	7	4.82		7	7.26		7	6.42		8	7.91
11	5	5.68		5	4.74		5	5.73		8	6.89

When we graphed these datasets, all datasets generate different kinds of plots which shows any regression algorithm are not able to interpret it



The statistical information for all these four datasets are approximately similar and can be computed as follows:

Anscombe's Data									
Observation	x1	y1		x2	y2	x3	y3	x4	y4
1	10	8.04		10	9.14	10	7.46	8	6.58
2	8	6.95		8	8.14	8	6.77	8	5.76
3	13	7.58		13	8.74	13	12.74	8	7.71
4	9	8.81		9	8.77	9	7.11	8	8.84
5	11	8.33		11	9.26	11	7.81	8	8.47
6	14	9.96		14	8.1	14	8.84	8	7.04
7	6	7.24		6	6.13	6	6.08	8	5.25
8	4	4.26		4	3.1	4	5.39	19	12.5
9	12	10.84		12	9.13	12	8.15	8	5.56
10	7	4.82		7	7.26	7	6.42	8	7.91
11	5	5.68		5	4.74	5	5.73	8	6.89
				Summary Statistics					
N	11	11		11	11	11	11	11	11
mean	9.00	7.50		9.00	7.500909	9.00	7.50	9.00	7.50
SD	3.16	1.94		3.16	1.94	3.16	1.94	3.16	1.94
r	0.82			0.82		0.82		0.82	

Hence, Anscombe's Quartet *says all the important features (including statistical properties) in the dataset must be visualised before implementing any machine learning algorithm, which will help to make a good fit model.*

References:

https://towardsdatascience.com/importance-of-data-visualization-anscombes-quartet-way-a325148b9fd2#:~:text=Anscombe's%20Quartet%20can%20be%20defined,when%20plotted%20on%20scatter%20plots.

Question: What is Pearson's R?

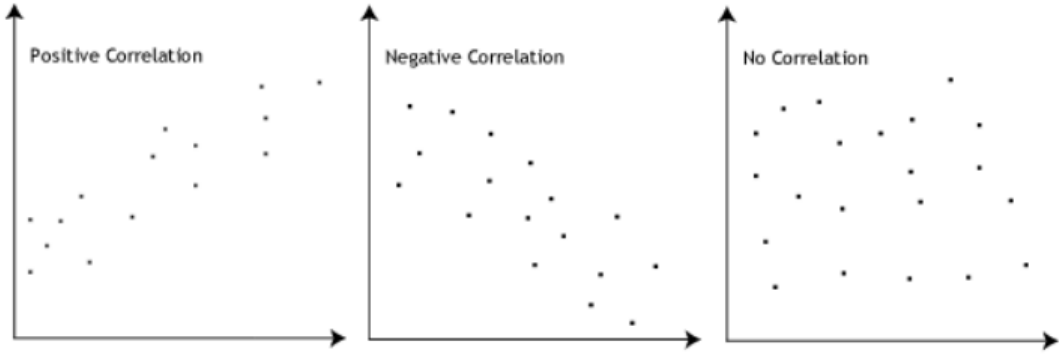
Answer:

In statistics, Pearson's correlation coefficient (**Pearson's r/ the Pearson product-moment correlation coefficient (PPMCC)/ bivariate correlation)** is a type of correlation which is used to measure the linear correlation between two continuous variables.

The Pearson's correlation coefficient (r) should be in range of [-1, +1] where:

- r = 1 indicates perfect positive correlation (i.e., both variables tend to change in the same direction)
- r = -1 indicates perfect negative correlation (i.e., both variables tend to change in different directions)
- r = 0 indicates there is no linear relation (both variables are independent)

Below are the graphs:



Formula for calculating Pearson's correlation coefficient is:

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

Here,

- r =correlation coefficient
- x_i =values of the x-variable in a sample
- $\bar{x}$ =mean of the values of the x-variable
- y_i =values of the y-variable in a sample
- $\bar{y}$ =mean of the values of the y-variable

Question: What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Answer:

Feature Scaling is a step of data pre-processing which is used to standardize the independent features present in the data in a fixed range.

Scaling is performed to handle highly varying magnitudes or values or units. In case if feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider

smaller values as the lower values, regardless of the unit of the values. It also helps in speeding up the calculations in an algorithm.
Note: scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

There are 2 types of scaling

1. Normalized scaling/ MinMax Scaling: rescales the values into a range of [0,1]

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

2. Standardized Scaling/ Z-score Normalization: rescales data to have a mean of 0 and a standard deviation of 1 (unit variance)

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

Normalized scaling vs Standardized scaling

Normalized Scaling	**Standardized Scaling**
Minimum and maximum value of features are used for scaling	Mean and standard deviation are used for scaling.
It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
It is really affected by outliers.	It is much less affected by outliers.
This transformation squishes the n-dimensional data into an n-dimensional unit hypercube.	It translates the data to the mean vector of original data to the origin and squishes or expands.
It is useful when we don't know about the distribution	It is useful when the feature distribution is Normal or Gaussian.

Reference link:
https://www.programsbuzz.com/interview-question/what-difference-between-normalized-scaling-and-standardized-scaling

---

Question: You might have observed that sometimes the value of VIF is infinite. Why does this happen?
Answer:
variance inflation Factor (VIF) is used to detect multicollinearity in the dataset. It calculated how well one independent variable is explained by all the other independent variables combined
Formula:

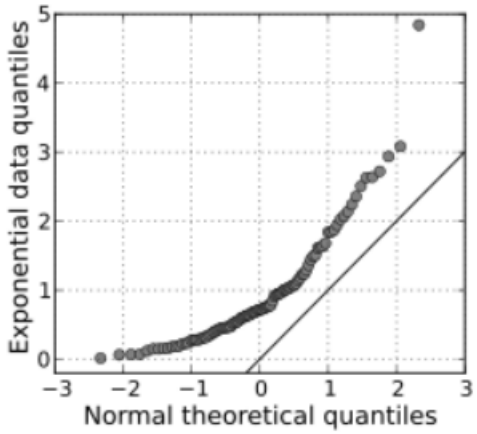
$$\text{VIF} = \frac{1}{1 - R^2}$$

In case of perfect correlation between 2 variables, we get R^2 is 1, which lead to VIF is infinity. To solve this issue, we need to drop one of variables from dataset to removing this perfect multicollinearity.

Question: 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) is a graphical tool or which is used to find out if two sets of data come from the common distribution, have common location and scale, have similar distribution shapes or have similar tail behaviour. *A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. For example, in case of median where 50% of the data lie above that point and 50% lie below it.*
Advantages:

- It can be used with sample sizes also
- Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.

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.



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.

References:
https://www.programsbuzz.com/interview-question/what-q-q-plot-explain-use-and-importance-q-q-plot-linear-regressio