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? (3 marks)

Answer:
Season: Maximum Bookings happened on Summer and Fall which gradually start decreasing from Winter to Spring
Month: the lowest booking happens in JAN which starts increases to max in July and Sept
weekday: there is no significance changes can be observed across the week days.
weathersit: Most of the bookings happen in Misty and Clear weather
holiday: on holidays the bookings were not like as expected
workingday: Close to 5000 booking happened on working day. This indicates, workingday can be a good predictor for the dependent variable

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

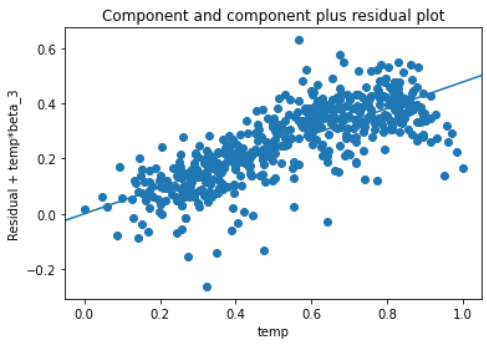
Answer:
it is to get n-1 dummies out of n categorical levels by removing the first level.

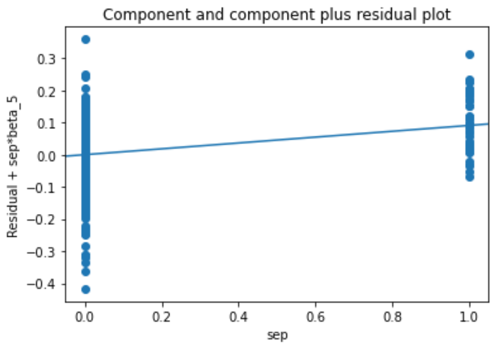
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

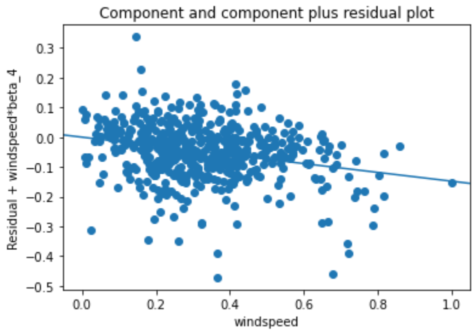
Answer:
Temp and atemp
temp and count
atemp and count

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

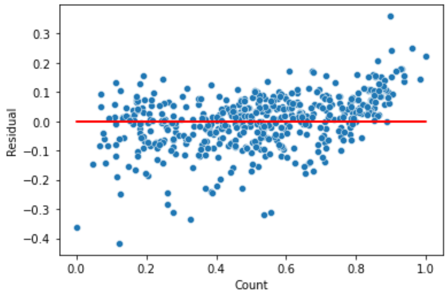
Answer:
We can check the linearity of the data by looking at the Residual vs Fitted plot. Ideally, this plot would not have a pattern where the red line (lowes smoother) is approximately horizontal at zero.
We can check this assumption using the Scale-Location plot.
In this plot we can see the fitted values vs the square root of the standardized residuals. Ideally, we would want to see the residual points equally spread around the red line, which would indicate constant variance.







Residual Errors Are Independent from Each Other and Predictors (x)



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

Answer:
Month
Weathershit
Season

General Subjective Questions

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

Answer:
Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting.

Different regression models differ based on – the kind of relationship between dependent and independent variables they are considering, and the number of independent variables getting used.

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.

While training the model we are given :
x: input training data (univariate – one input variable(parameter))
y: labels to data (supervised learning)

When training the model – it fits the best line to predict the value of y for a given value of x. The model gets the best regression fit line by finding the best θ1 and θ2 values.
θ1: intercept
θ2: coefficient of x
Once we find the best θ1 and θ2 values, we get the best fit line. So when we are finally using our model for prediction, it will predict the value of y for the input value of x.

2. Explain the Anscombe’s quartet in detail. (3 marks)

Answer:
Anscombe’s quartet comprises four data sets that have nearly identical simple descriptive statistics yet have very different distributions and appear very different when graphed.

Anscombe’s Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots.

This tells us about the importance of visualizing the data before applying various algorithms out there to build models out of them which suggests that the data features must be plotted to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data, etc. Also, the Linear Regression can only be considered a fit for the data with linear relationships and is incapable of handling any other kind of datasets.

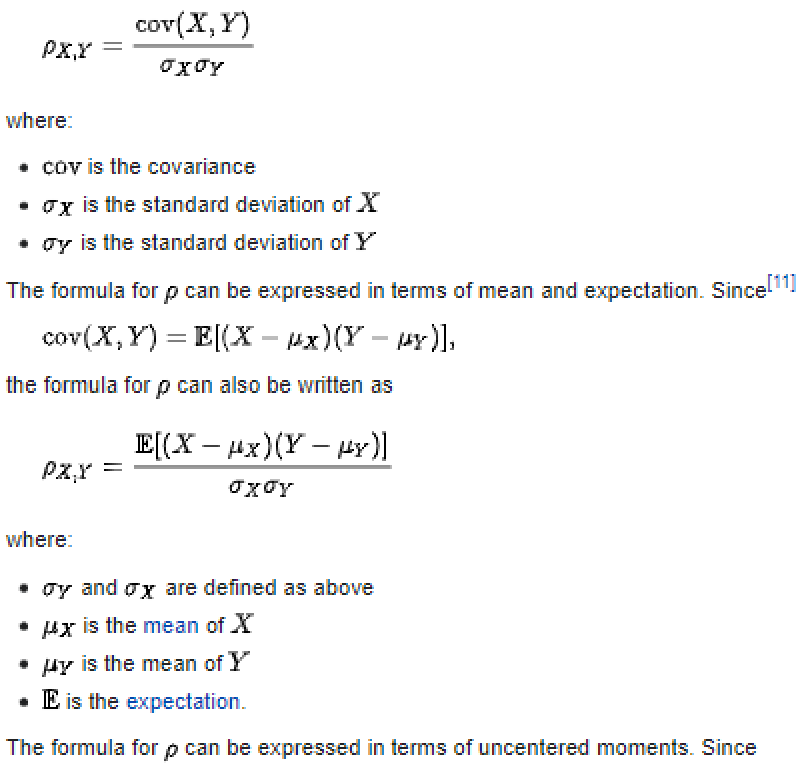
The four datasets can be described as:

Dataset 1: this fits the linear regression model well.
Dataset 2: this could not fit linear regression model on the data quite well as the data is non-linear.

Dataset 3: shows the outliers involved in the dataset which cannot be handled by linear regression model
Dataset 4: shows the outliers involved in the dataset which cannot be handled by linear regression model

3. What is Pearson’s R? (3 marks)

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, when applied to a population, is commonly represented by the Greek letter ρ (rho) and may be referred to as the population correlation coefficient or the population Pearson correlation coefficient. Given a pair of random variables.



$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y}$$

where:

- cov is the covariance
- σ_X is the standard deviation of X
- σ_Y is the standard deviation of Y

The formula for ρ can be expressed in terms of mean and expectation. Since[11]

$$\operatorname{cov}(X,Y) = \mathbb{E}[(X-\mu_X)(Y-\mu_Y)],$$

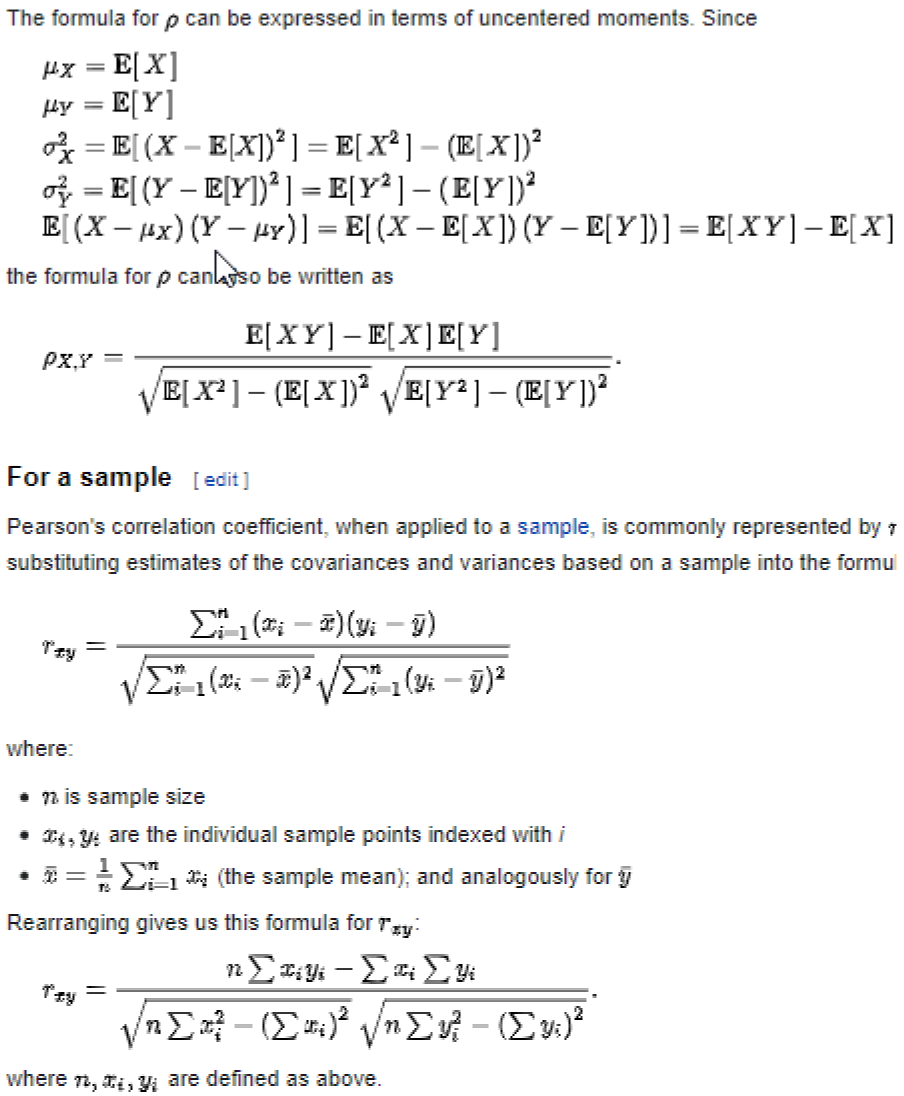
the formula for ρ can also be written as

$$\rho_{X,Y} = \frac{\mathbb{E}[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$$

where:

- σ_Y and σ_X are defined as above
- μ_X is the mean of X
- μ_Y is the mean of Y
- $\mathbb{E}$ is the expectation.

The formula for ρ can be expressed in terms of uncentered moments. Since



The formula for ρ can be expressed in terms of uncentered moments. Since

$$\begin{aligned}
\mu_X &= \mathbb{E}[X] \\
\mu_Y &= \mathbb{E}[Y] \\
\sigma_X^2 &= \mathbb{E}[(X-\mathbb{E}[X])^2] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 \\
\sigma_Y^2 &= \mathbb{E}[(Y-\mathbb{E}[Y])^2] = \mathbb{E}[Y^2] - (\mathbb{E}[Y])^2 \\
\mathbb{E}[(X-\mu_X)(Y-\mu_Y)] &= \mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])] = \mathbb{E}[XY] - \mathbb{E}[X]
\end{aligned}$$

the formula for ρ can also be written as

$$\rho_{X,Y} = \frac{\mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y]}{\sqrt{\mathbb{E}[X^2] - (\mathbb{E}[X])^2}\,\sqrt{\mathbb{E}[Y^2] - (\mathbb{E}[Y])^2}}.$$

For a sample [edit]

Pearson's correlation coefficient, when applied to a sample, is commonly represented by r substituting estimates of the covariances and variances based on a sample into the formul

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

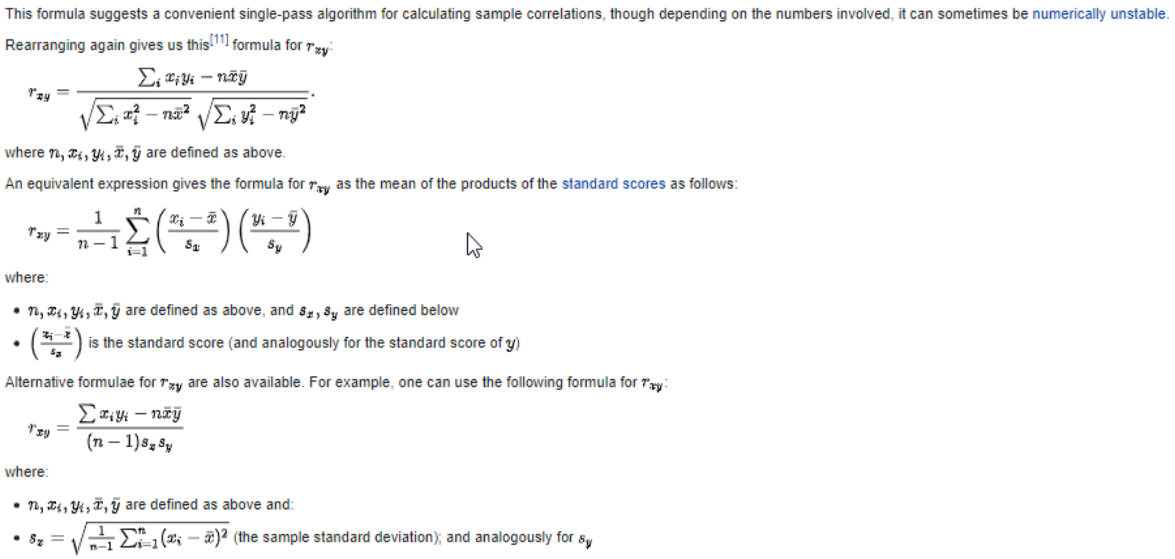
where:

- n is sample size
- x_i, y_i are the individual sample points indexed with *i*
- $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ (the sample mean); and analogously for $\bar{y}$

Rearranging gives us this formula for r_{xy}:

$$r_{xy} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\,\sqrt{n\sum y_i^2 - (\sum y_i)^2}}.$$

where n, x_i, y_i are defined as above.



This formula suggests a convenient single-pass algorithm for calculating sample correlations, though depending on the numbers involved, it can sometimes be numerically unstable.

Rearranging again gives us this[11] formula for r_{xy}:

$$r_{xy} = \frac{\sum_i x_i y_i - n\bar{x}\bar{y}}{\sqrt{\sum_i x_i^2 - n\bar{x}^2}\,\sqrt{\sum_i y_i^2 - n\bar{y}^2}}.$$

where $n, x_i, y_i, \bar{x}, \bar{y}$ are defined as above.

An equivalent expression gives the formula for r_{xy} as the mean of the products of the standard scores as follows:

$$r_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

where:

- $n, x_i, y_i, \bar{x}, \bar{y}$ are defined as above, and s_x, s_y are defined below
- $\left(\frac{x_i - \bar{x}}{s_x}\right)$ is the standard score (and analogously for the standard score of y)

Alternative formulae for r_{xy} are also available. For example, one can use the following formula for r_{xy}:

$$r_{xy} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{(n-1)s_x s_y}$$

where:

- $n, x_i, y_i, \bar{x}, \bar{y}$ are defined as above and:
- $s_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$ (the sample standard deviation); and analogously for s_y

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

Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units. 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.

In the machine learning algorithms if the values of the features are closer to each other there are chances for the algorithm to get trained well and faster instead of the data set where the data points or features values have high differences with each other will take more time to understand the data and the accuracy will be lower.
So if the data in any conditions has data points far from each other, scaling is a technique to make them closer to each other or in simpler words, we can say that the scaling is used for making data points generalized so that the distance between them will be lower.

Normalising typically means to transform your observations xx into f(x)f(x) (where ff is a measurable, typically continuous, function) such that they look normally distributed. Some examples of transformations for normalising data are power transformations.

Scaling simply means f(x)=cxf(x)=cx, c∈Rc∈R, this is, multiplying your observations by a constant cc which changes the scale (for example from nanometers to kilometers).

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

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. (3 marks)

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.

Normal Distribution:

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

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.

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.