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?

Ans:

Categorical variables can have a significant impact on the dependent variable in a regression model in order to include categorical variables in a regression model, they are typically converted into dummy variables (also known as indicator variables). This involves creating a separate binary variable for each category of the variable.

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.1925	0.029	6.625	0.000	0.135	0.250
yr	0.2304	0.008	28.850	0.000	0.215	0.246
holiday	-0.0515	0.027	-1.923	0.055	-0.104	0.001
workingday	0.0434	0.011	3.784	0.000	0.021	0.066
temp	0.4872	0.034	14.176	0.000	0.420	0.555
hum	-0.1663	0.037	-4.439	0.000	-0.240	-0.093
windspeed	-0.1849	0.025	-7.268	0.000	-0.235	-0.135
season_2	0.1045	0.019	5.481	0.000	0.067	0.142
season_3	0.0430	0.023	1.894	0.059	-0.002	0.088
season_4	0.1531	0.014	10.985	0.000	0.126	0.181
mnth_3	0.0352	0.015	2.309	0.021	0.005	0.065
mnth_4	0.0192	0.021	0.912	0.362	-0.022	0.061
mnth_5	0.0399	0.020	1.951	0.052	-0.000	0.080
mnth_8	0.0477	0.017	2.773	0.006	0.014	0.082
mnth_9	0.1179	0.017	6.929	0.000	0.084	0.151
mnth_10	0.0457	0.017	2.654	0.008	0.012	0.079
weekday_Sunday	0.0531	0.014	3.672	0.000	0.025	0.081
weathersit_2	-0.0594	0.010	-5.747	0.000	-0.080	-0.039
weathersit_3	-0.2531	0.026	-9.697	0.000	-0.304	-0.202

Example: spring, winter fall under season category and weathersir_12, weathersit_3 under weathersit category are encoded with dummy variables.

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

Ans:

The drop_first=True option in dummy variable creation is used to avoid the problem of multicollinearity in regression analysis.

Multicollinearity is a situation where two or more variables in a regression model are highly correlated, meaning one can be linearly predicted from the others with a substantial degree of accuracy. This can lead to unstable and unreliable estimates of the model parameters.

When you create dummy variables for a categorical variable with n categories, you actually create n new variables, each representing one category. Each of these variables takes the value 1 if the category is present, and 0 otherwise.

However, if you include all n dummy variables in your model, you introduce multicollinearity because one variable can be perfectly predicted from the others. For example, if you have three categories A, B, and C, and you know that categories A and B are 0, then category C must be 1.

To avoid this, you can use the drop_first=True option, which will create n-1 dummy variables, effectively dropping the first category. This serves as the reference category against which the effects of the other categories are measured, and it eliminates the problem of multicollinearity.

So, it's important to use drop_first=True to ensure the validity and reliability of your regression model.

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

Ans:

Therefore, the pair plot displays the strongest correlation for the registered variable (correlation 0.945) prior to model creation and training. However, we are not training the model with random and registered data

from our pre-processed training set. Registered + casual = CNT. This might cause the model to become overfit and leak important information.

After removing these two factors, the target variable, cnt, has the lowest association with atemp, which is followed by temp.

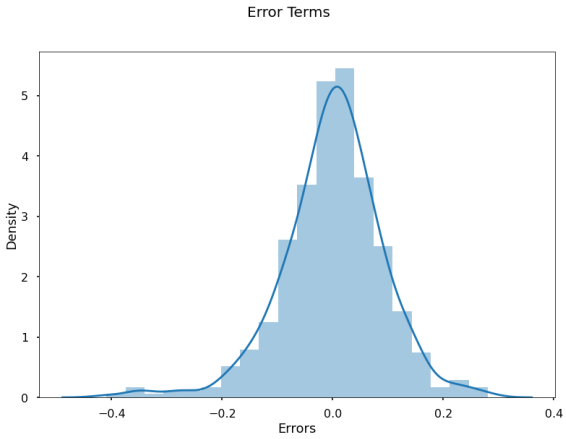
The correlation coefficient between atemp and cnt is 0.631, according to the correlation heatmap. Additionally, there is 0.627 link between temp and cnt.

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

Ans:

To validate assumptions of the model, and hence the reliability for inference, we go with the following procedures:

- Residual Analysis: We need to check if the error terms are also normally distributed (which is in fact, one of the major assumptions of linear regression). I have plotted the histogram of the error terms, and this is what it looks like:



- Error terms are independent of each other: Handled properly in the model. The predictor variables are independent of each other. Multicollinearity issue is not there because VIF (Variance Inflation Factor) for all predictor variables are below 5.

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

Ans:

1) temp (coef: 0.554)

2) yr (coef: 0.231)

3) month_9 (coef: 0.09)

General Subjective Questions:

1) Explain the linear regression algorithm in detail.

Linear regression is a statistical method that allows us to study relationships between two continuous (quantitative) variables:

One variable, denoted x, is regarded as the predictor, explanatory, or independent variable.The other variable, denoted y, is regarded as the response, outcome, or dependent variable.

The goal of linear regression is to model the expected value of y given the value of x.

Here's the basic form of a linear regression model:

$$y=\beta 0+\beta 1x+\epsilon$$

Where,

y is the dependent variable.

x is the independent variable.

β0 is the y-intercept.

β1 is the slope.

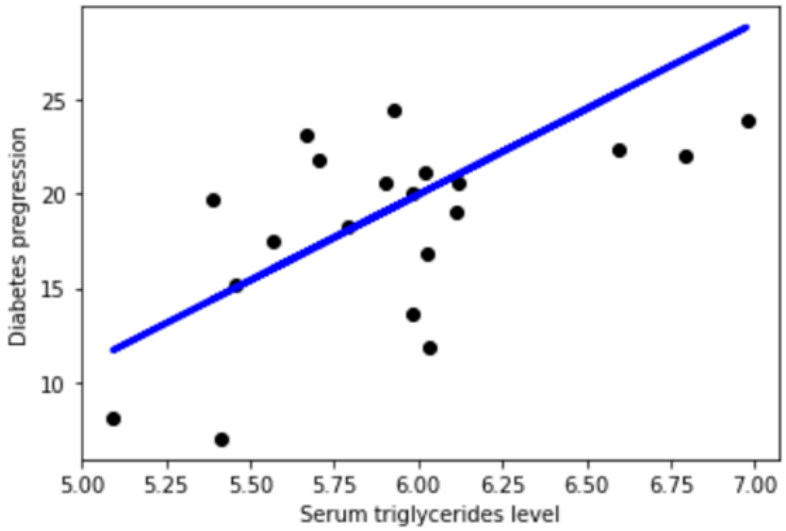
ϵ is the error term (also known as the residual).

The β values are called the model coefficients:

These values are "learned" during the model fitting process using the "least squares" criterion.

Then, the fitted model can be used to make predictions!

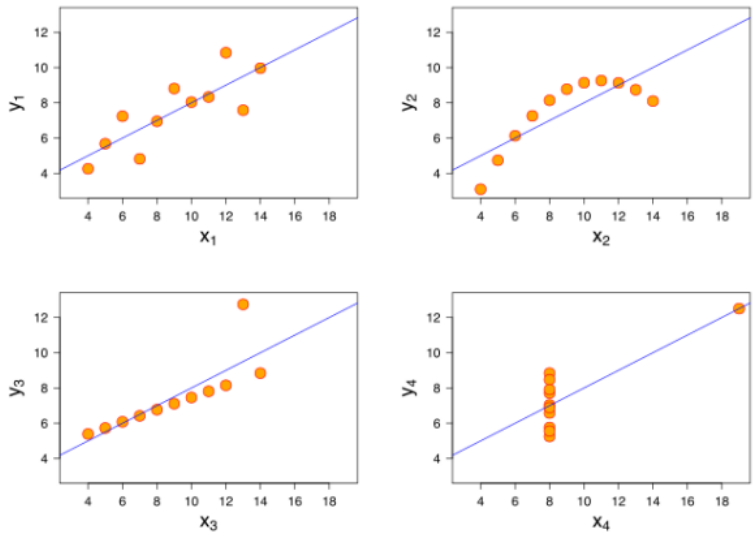
The least squares method minimizes the sum of the squared residuals, i.e., the difference between the observed and predicted values, in the dataset.



2) Explain the Anscombe's quartet in detail.

Ans:

The four data sets that make up Anscombe's quartet contain virtually similar basic descriptive statistics, yet their distributions and visual representations on a graph are drastically different.



All four sets are identical when examined using simple summary statistics, but vary considerably when graphed.

1) The first scatter plot (top left) appears to be a simple linear relationship, corresponding to two variables correlated where y could be modelled as gaussian with mean linearly dependent on x.

2) The second graph (top right) is not distributed normally; while a relationship between the two variables is obvious, it is not linear, and the Pearson correlation coefficient is not relevant. A more general regression and the corresponding coefficient of determination would be more appropriate.

3) In the third graph (bottom left), the distribution is linear, but should have a different regression line (a robust regression would have been called for). The calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.

4) the fourth graph (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

3) What is Pearson's R?

A measure of the linear connection between two sets of data is Pearson's R, often known as the correlation coefficient. It is effectively a normalized measurement of covariance, with a value that is always between −1 and 1. It is defined as the ratio between the covariance of two variables and the product of their standard deviations. Similar to covariance, this measure only captures a linear connection between variables; it misses numerous other kinds of correlations or relationships.

1) A correlation coefficient of 1 means that for every positive increase in one variable, there is a positive increase of a fixed proportion in the other. For example, shoe sizes go up in (almost) perfect correlation with foot length.

2) A correlation coefficient of -1 means that for every positive increase in one variable, there is a negative decrease of a fixed proportion in the other. For example, the amount of gas in a tank decrease in (almost) perfect correlation with speed.

3) Zero means that for every increase, there isn't a positive or negative increase. The two just aren't related.

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

Scaling is a preprocessing step in machine learning that transforms the range of values of features in your dataset. It's performed to ensure that all features contribute equally to the model, regardless of their original scales. This is important because many machine learning algorithms are sensitive to the scale of the features.

Here's a brief explanation of the two common types of scaling:

1) Normalization: This method scales the data to a fixed range, usually 0 to 1. The formula for normalization, also known as min-max scaling, is:

$$x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

where xnew is the normalized value, x is the original value, and xmin and xmax are the minimum and maximum values of the feature, respectively. Normalization is useful when you know the bounds of your data, and when your data has a uniform distribution.

2) Standardization: This method scales the data to have a mean of 0 and a standard deviation of 1. The formula for standardization, also known as z-score normalization, is:

$$x_{new} = \frac{x - \mu}{\sigma}$$

where xnew is the standardized value, x is the original value, μ is the mean of the feature, and σ is the standard deviation of the feature. Standardization is useful when your data follows a Gaussian distribution, and it's less affected by outliers compared to normalization.

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

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).

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

A Q-Q plot, or quantile-quantile plot, is a graphical tool used to assess if a dataset follows a particular theoretical distribution. It's a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight.

Here's how it works:

- One axis, say the x-axis, represents the quantiles from the theoretical distribution.

- The other axis, say the y-axis, represents the quantiles from the dataset.
- Each point on the plot corresponds to a specific data point in the dataset, and its coordinates are the values of the theoretical and sample quantiles.

In the context of linear regression, Q-Q plots are used to check the assumption of normality of the errors (residuals). This is an important assumption in linear regression. If the residuals are normally distributed, it lends more credibility to the results of the regression analysis.

Here's how you interpret a Q-Q plot in this context:

- If the points lie along or close to the line y=x, this suggests that the residuals are normally distributed, which is a good sign.
- If the points deviate from the line in a systematic way, this suggests that the residuals are not normally distributed, which could be a problem. For example, if the points bend upwards away from the line at the ends, this suggests the residuals have heavier tails than a normal distribution.

In summary, a Q-Q plot is a handy tool for checking the normality of residuals in a linear regression model, which is a key assumption of this type of model. If the assumption is violated, the model's predictions and inference may not be reliable.