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:

   The final Multiple Linear Regression model contains many predictor variables that are categorical in nature and some of them have been encoded to dummy variables.

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.2574	0.025	10.261	0.000	0.208	0.307
yr	0.2354	0.008	28.110	0.000	0.219	0.252
workingday	0.0233	0.009	2.611	0.009	0.006	0.041
temp	0.3951	0.030	13.272	0.000	0.337	0.454
windspeed	-0.1430	0.025	-5.609	0.000	-0.193	-0.093
spring	-0.1296	0.016	-8.287	0.000	-0.160	-0.099
winter	0.0315	0.015	2.133	0.033	0.002	0.060
weathersit_2	-0.0792	0.009	-8.890	0.000	-0.097	-0.062
weathersit_3	-0.2976	0.025	-11.744	0.000	-0.347	-0.248
month_10	0.0461	0.018	2.549	0.011	0.011	0.082
month_3	0.0425	0.015	2.865	0.004	0.013	0.072
month_9	0.0734	0.016	4.629	0.000	0.042	0.105

Omnibus:	78.141	Durbin-Watson:	2.000
Prob(Omnibus):	0.000	Jarque-Bera (JB):	219.721
Skew:	-0.741	Prob(JB):	1.94e-48
Kurtosis:	5.853	Cond. No.	15.8

   spring, winter falls under season category and have been dummy encoded. weathersit_2 and weathersit_3 falls under weathersit category and have been dummy encoded. Similarly, month variables fall under mnth category and have been dummy encoded. We can infer from above image that these variables are statistically significant and explain the variance in model very well.

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

   Answer:

   Reason 1:

   To avoid dummy variable trap. Dummy variable trap might lead to multicollinearity issue among dummy variables. This may lead to violation of assumptions of linear

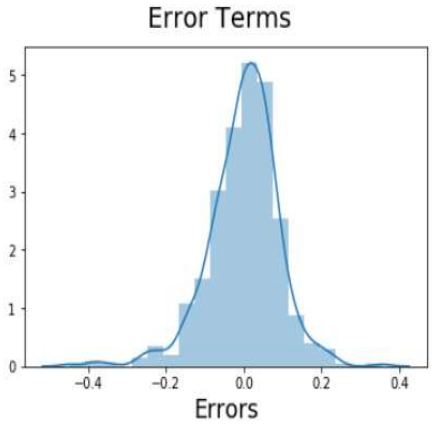
regression. So, if we have k levels where k >=3, we only use k-1 levels while dummy variable encoding. The dropped level gets handled by intercept as a base case.

Reason 2:
Suppose there is category of colors like Red, Green, and Blue. This category belongs to nominal categorical variable. There is no order or relation among Red, Green, Blue. We cannot simply label encode them as 1, 2, 3 because this will confuse our model which might lead to order-based bias like Red < Green < Blue. To avoid this, we dummy encode cases like this. Also, model cannot understand string/text data therefore it is necessary to convert them into numbers.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)
Answer:
*So, before model building and training, the pair plot shows highest correlation for registered variable having correlation 0.945. But we are not using casual and registered in our pre-processed training data for model training**.** casual + registered = cnt. This might leak out the crucial information and model might get overfit.*

*So, excluding these two variables **atemp** is having highest correlation with target variable **cnt** which is followed by **temp**.*

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



The residuals are following the normally distribution with a mean 0. All good!

- ***Linear relationship between predictor variables and target variable:***
  *This is happening because all the predictor variables are statistically significant (p-values are less than **0.05**). Also, R-Squared value on training set is **0.830** and adjusted R-Squared value on training set is **0.826**. This means that variance in data is being explained by all these predictor variables.*

- ***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? (2 marks)
   Answer:
   Top 3 features significantly contributing towards demand of shared bikes are:
   1) ***temp** (coef: **0.3951**)*
   2) ***yr** (coef: **0.2354**)*
   3) ***month_9** (coef: **0.0734**)*

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)
   Answer:
   Linear regression is a data analysis technique that predicts the value of unknown data by using another related and known data value. It mathematically models the unknown or dependent variable and the known or independent variable as a linear equation. For instance, suppose that you have data about your expenses and income for last year. Linear regression techniques analyze this data and determine that your expenses are half your income. They then calculate an unknown future expense by halving a future known income.

   Why it is important ?
   Linear regression models are relatively simple and provide an easy-to-interpret mathematical formula to generate predictions. Linear regression is an established statistical technique and applies easily to software and computing. Businesses use it to reliably and predictably convert raw data into business intelligence and actionable insights. Scientists in many fields, including biology and the behavioral, environmental, and social sciences, use linear regression to conduct preliminary data analysis and predict future trends. Many data science methods, such as machine learning and artificial intelligence, use linear regression to solve complex problems.

How does linear regression work?

At its core, a simple linear regression technique attempts to plot a line graph between two data variables, x and y. As the independent variable, x is plotted along the horizontal axis. Independent variables are also called explanatory variables or predictor variables. The dependent variable, y, is plotted on the vertical axis. You can also refer to y values as response variables or predicted variables.

Steps in linear regression

*For this overview, consider the simplest form of the line graph equation between y and x; y=c*x+m, where c and m are constant for all possible values of x and y. So, for example, suppose that the input dataset for (x,y) was (1,5), (2,8), and (3,11). To identify the linear regression method, you would take the following steps:*

1. *Plot a straight line, and measure the correlation between 1 and 5.*
2. *Keep changing the direction of the straight line for new values (2,8) and (3,11) until all values fit.*
3. *Identify the linear regression equation as y=3*x+2.*
4. *Extrapolate or predict that y is 14 when x is*

What is linear regression in machine learning?

In machine learning, computer programs called algorithms analyze large datasets and work backward from that data to calculate the linear regression equation. Data scientists first train the algorithm on known or labeled datasets and then use the algorithm to predict unknown values. Real-life data is more complicated than the previous example. That is why linear regression analysis must mathematically modify or transform the data values to meet the following four assumptions.

Linear relationship

A linear relationship must exist between the independent and dependent variables. To determine this relationship, data scientists create a scatter plot—a random collection of x and y values—to see whether they fall along a straight line. If not, you can apply nonlinear functions such as square root or log to mathematically create the linear relationship between the two variables.

Residual independence

Data scientists use residuals to measure prediction accuracy. A residual is the difference between the observed data and the predicted value. Residuals must not have an identifiable pattern between them. For example, you don't want the residuals to grow larger with time. You can use different mathematical tests, like the Durbin-Watson test, to determine residual independence. You can use dummy data to replace any data variation, such as seasonal data.

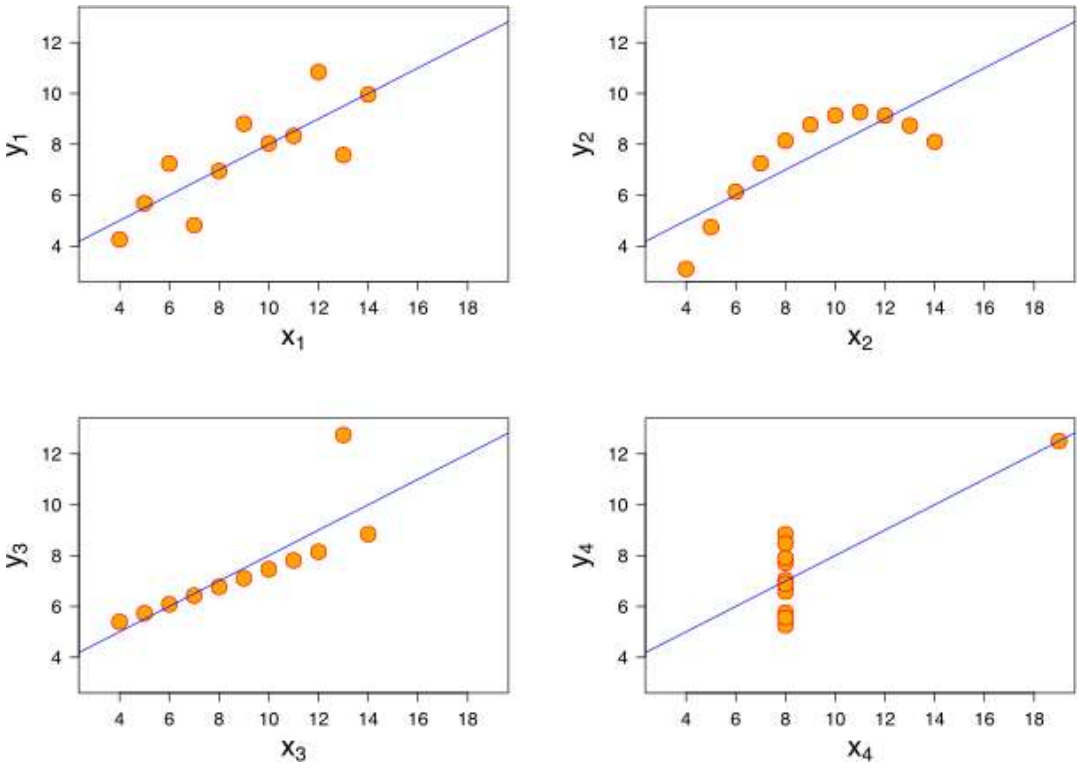
Normality

Graphing techniques like Q-Q plots determine whether the residuals are normally distributed. The residuals should fall along a diagonal line in the center of the graph. If the residuals are not normalized, you can test the data for random outliers or values that are not typical. Removing the outliers or performing nonlinear transformations can fix the issue.

Homoscedasticity

Homoscedasticity assumes that residuals have a constant variance or standard deviation from the mean for every value of x. If not, the results of the analysis might not be accurate. If this assumption is not met, you might have to change the dependent variable. Because variance occurs naturally in large datasets, it makes sense to change the scale of the dependent variable. For example, instead of using the population size to predict the number of fire stations in a city, might use population size to predict the number of fire stations per person.

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.



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.*

Property	Value	Accuracy
Mean of x	9	exact
Sample variance of x : s_x^2	11	exact
Mean of y	7.50	to 2 decimal places
Sample variance of y : s_y^2	4.125	±0.003
Correlation between x and y	0.816	to 3 decimal places
Linear regression line	$y = 3.00 + 0.500x$	to 2 and 3 decimal places, respectively
Coefficient of determination of the linear regression : R^2	0.67	to 2 decimal places

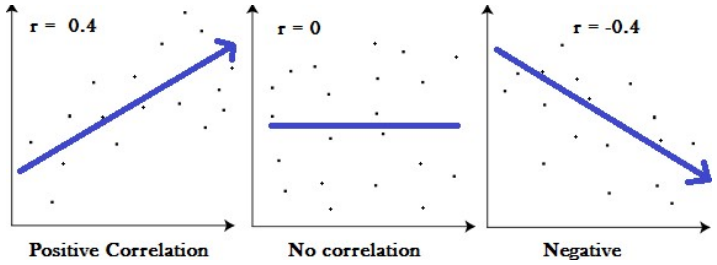
Anscombe's quartet

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

3. What is Pearson's R? (3 marks)

Answer:

Pearson's R or correlation coefficient is a measure of linear correlation between two sets of data. It is the ratio between the covariance of two variables and 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. As with covariance itself, the measure can only reflect a linear correlation of variables, and ignores many other types of relationship or correlation.



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.*

The absolute value of the correlation coefficient gives us the relationship strength. The larger the number, the stronger the relationship. For example, |-.95| = .95, which has a stronger relationship than .55.

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

Answer:

Scaling is a method used to normalize the range of independent variables or features of data.

Since the range of values of raw data varies widely, in some machine learning algorithms, objective functions will not work properly without normalization. For example, many classifiers calculate the distance between two points by the Euclidean distance. If one of the features has a broad range of values, the distance will be governed by this particular feature. Therefore, the range of all features should be normalized so that each feature contributes approximately proportionately to the final distance.

Another reason why feature scaling is applied is that gradient descent converges much faster with feature scaling than without it.

Normalization:

Also known as min-max scaling or min-max normalization, it is the simplest method and consists of rescaling the range of features to scale the range in [0, 1]. The general formula for normalization is given as:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Here, max(x) and min(x) are the maximum and the minimum values of the feature respectively.

Standardization:

Feature standardization makes the values of each feature in the data have zero mean and unit variance. The general method of calculation is to determine the distribution mean and standard deviation for each feature and calculate the new data point by the following formula:

$$x' = \frac{x - \bar{x}}{\sigma}$$

Here, σ is the standard deviation of the feature vector, and $\bar{x}$ is the average of the feature vector.

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

   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 (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. (3 marks)

   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.

   This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.

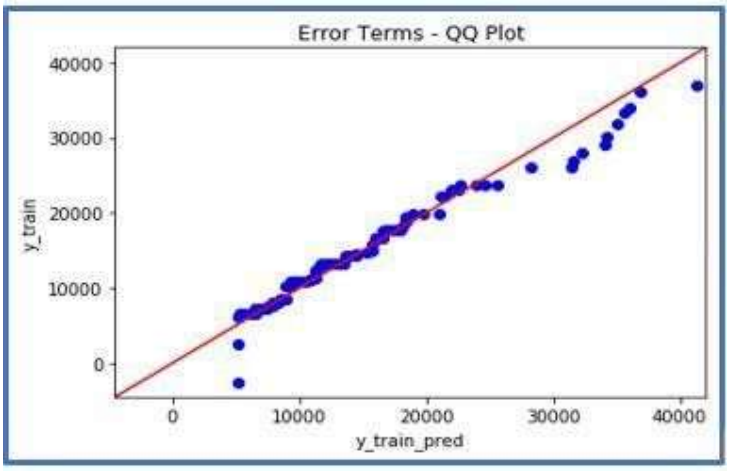
   Interpretation:

   A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.

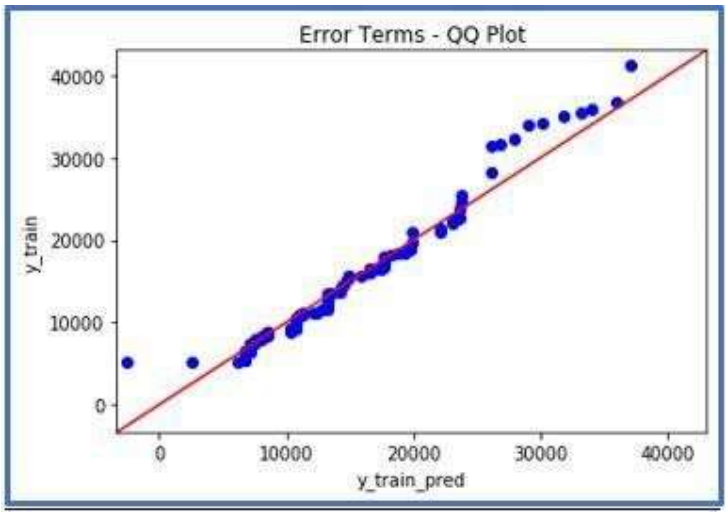
   Below are the possible interpretations for two data sets.

   1) *Similar distribution: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x -axis*

2) *Y-values < X-values: If y-quantiles are lower than the x-quantiles.*



3) *X-values < Y-values: If x-quantiles are lower than the y-quantiles.*



4) *Different distribution: If all point of quantiles lies away from the straight line at an angle of 45 degree from x -axis*

statsmodels.api *provide* ***qqplot*** *and* ***qqplot_2samples*** *to plot* ***Q-Q*** *graph for single and two different data sets respectively.*