Bike Sharing Assignment Questions

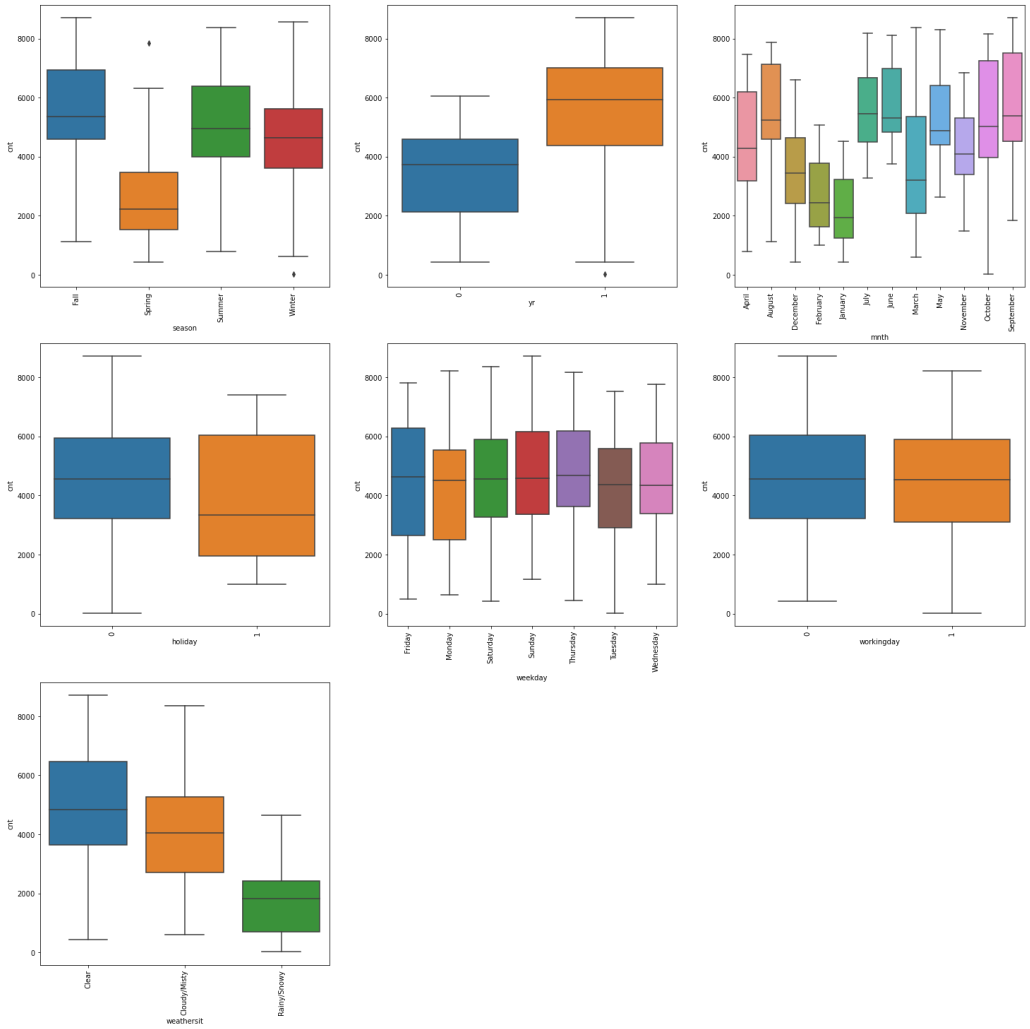
Assignment-based Subjective Questions

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

   Ans: The relationship between the dependent variable (cnt) and the categorical variables are observed to be as follows:

   - The demand is highest during fall and least during spring
   - The bike demand has increased considerably in 2019, as compared to 2018
   - No bikes are rented during weather corresponding to category 4 (Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog)
   - The median number of bikes rented is more on weekdays than on weekends - stating that the prime target population of the bike renting is the working professionals

   The corresponding box-plot distribution is given as (zoom to enlarge) –



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

Ans:

- 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.
- For the below example – we have a category called “Furnishing Status” that has values corresponding to Unfurnished, Semi-Furnished and Furnished

Value	Indicator Variable	
Furnishing Status	furnished	semi-furnished
furnished	1	0
semi-furnished	0	1
unfurnished	0	0

- We see from the above example that the furnishing status of all categories can be explained as –

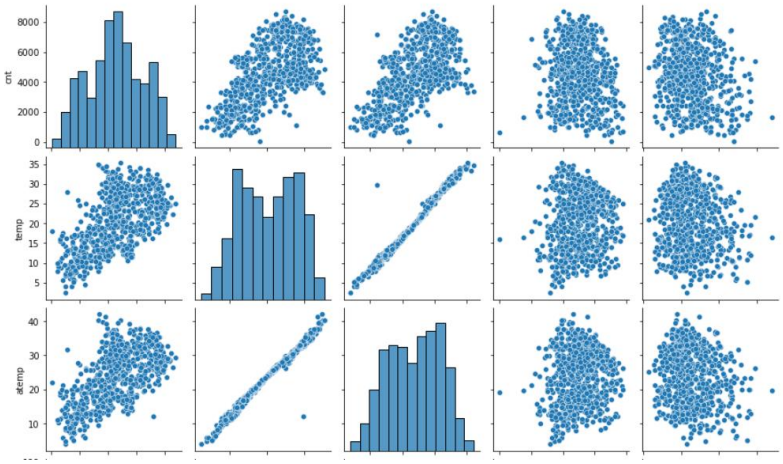
  Furnished – 10
  Semi-furnished – 01
  Unfurnished – 00

- Thus, for a category with n-values, we require only n-1 dummy variables

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

Ans:

- Looking at the pair-plot, we see both ‘temp’ and ‘atemp’ have very high correlation with the target variable (cnt):



- These two variables also have a very high correlation between them, and later using VIF we find that using temp is more favourable than using atemp since temp has a lower VIF, leading us to eventually drop atemp.
- So, from the above understanding, we can state that temp has the highest correlation with the target variable.

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

Ans:

- The linear regression model is validated using the following parameters obtained from the OLS summary:

OLS Regression Results

Dep. Variable:	cnt	R-squared:	0.829
Model:	OLS	Adj. R-squared:	0.825
Method:	Least Squares	F-statistic:	200.6
Date:	Wed, 08 Dec 2021	Prob (F-statistic):	9.63e-182
Time:	14:16:09	Log-Likelihood:	467.99
No. Observations:	510	AIC:	-910.0
Df Residuals:	497	BIC:	-854.9
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.1129	0.025	4.474	0.000	0.063	0.162
yr	0.2414	0.009	27.639	0.000	0.224	0.259
holiday	-0.0709	0.027	-2.651	0.008	-0.123	-0.018
season_Spring	-0.1590	0.017	-9.102	0.000	-0.193	-0.125
season_Winter	0.0940	0.015	6.441	0.000	0.065	0.123
mnth_December	-0.0605	0.019	-3.248	0.001	-0.097	-0.024
mnth_July	-0.0505	0.017	-2.899	0.004	-0.085	-0.016
mnth_March	0.0569	0.019	3.032	0.003	0.020	0.094
mnth_November	-0.0825	0.020	-4.188	0.000	-0.121	-0.044
mnth_September	0.0519	0.016	3.236	0.001	0.020	0.083
weathersit_Clear	0.0844	0.009	9.016	0.000	0.066	0.103
weathersit_Rainy/Snowy	-0.2027	0.027	-7.611	0.000	-0.255	-0.150
temp	0.4287	0.035	12.242	0.000	0.360	0.498

Omnibus:	76.146	Durbin-Watson:	2.093
Prob(Omnibus):	0.000	Jarque-Bera (JB):	169.699
Skew:	-0.801	Prob(JB):	1.41e-37
Kurtosis:	5.328	Cond. No.	15.7

a. R-Squared: 0.829. Meaning that 82.9% of the variance in Bike Counts is explained by the model. This is a decent R-squared value.

b. Adjusted R-Squared: 0.825. Since Adjusted R-Squared is not significantly less than R-Squared, this implies that the independent variables taken into consideration are all significant.

c. P-value: The P-values of all the independent variables individually is < 0.05. This further reinforces the above conclusion that all independent variables are individually significant.

d. F-Statistics: F statistic has a very low p-value (practically 0). Meaning that the model fit is statistically significant, and the explained variance isn't purely by chance.

These are the primary parameters that were evaluated to finalize on the model.

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

Ans:

Based on the values of the coefficients from the stats model summary (below)

OLS Regression Results

Dep. Variable:	cnt	R-squared:	0.829
Model:	OLS	Adj. R-squared:	0.825
Method:	Least Squares	F-statistic:	200.6
Date:	Wed, 08 Dec 2021	Prob (F-statistic):	9.63e-182
Time:	14:16:09	Log-Likelihood:	467.99
No. Observations:	510	AIC:	-910.0
Df Residuals:	497	BIC:	-854.9
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.1129	0.025	4.474	0.000	0.063	0.162
yr	0.2414	0.009	27.639	0.000	0.224	0.259
holiday	-0.0709	0.027	-2.651	0.008	-0.123	-0.018
season_Spring	-0.1590	0.017	-9.102	0.000	-0.193	-0.125
season_Winter	0.0940	0.015	6.441	0.000	0.065	0.123
mnth_December	-0.0605	0.019	-3.248	0.001	-0.097	-0.024
mnth_July	-0.0505	0.017	-2.899	0.004	-0.085	-0.016
mnth_March	0.0569	0.019	3.032	0.003	0.020	0.094
mnth_November	-0.0825	0.020	-4.188	0.000	-0.121	-0.044
mnth_September	0.0519	0.016	3.236	0.001	0.020	0.083
weathersit_Clear	0.0844	0.009	9.016	0.000	0.066	0.103
weathersit_Rainy/Snowy	-0.2027	0.027	-7.611	0.000	-0.255	-0.150
temp	0.4287	0.035	12.242	0.000	0.360	0.498

Omnibus:	76.146	Durbin-Watson:	2.093
Prob(Omnibus):	0.000	Jarque-Bera (JB):	169.699
Skew:	-0.801	Prob(JB):	1.41e-37
Kurtosis:	5.328	Cond. No.	15.7

The top-3 features are:

- Temperature: coeff 0.4287
- Year: coeff 0.2414
- Season_Winter (season = 4): coeff 0.0940

Assignment-based Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans:

- Linear Regression is the supervised Machine Learning model in which the model finds the best fit linear line between the independent and dependent variable
- That is - it finds the linear relationship between the dependent(y) and independent variable(x).

- Linear Regression is of two types: Simple and Multiple.

- Simple Linear Regression is where only one independent variable is present and the model is expected to find the linear relationship of it with the dependent variable
- Equation of Simple Linear Regression, where b0 is the intercept, b1 is coefficient or slope, x is the independent variable and y is the dependent variable.

$$y = b_o + b_1 x$$

- Multiple Linear Regression there are more than one independent variables for the model to find the relationship.
- Equation of Multiple Linear Regression, where b0 is the intercept, b1, b2, b3, b4…, bn are coefficients or slopes of the independent variables x1, x2, x3,….,xn and y is the dependent variable.

$$y = b_o + b_1 x_1 + b_2 x_2 + b_3 x_3 \ldots . + b_n x_n$$

- A Linear Regression model's main aim is to find the best fit linear line and the optimal values of intercept and coefficients such that the error is minimized.

- Linear Regression Algorithm considers the following assumptions –

    1. **Linearity**: It states that the dependent variable Y should be linearly related to independent variables. This assumption can be checked by plotting a scatter plot between both variables
    2. **Homoscedasticity**: The variance of the error terms should be constant
    3. **Independence/No Multicollinearity**: The variables should be independent of each other
    4. The error terms should be **normally distributed**
    5. **No Autocorrelation**: The error terms should be independent of each other

- Linear Regression Algorithm can be summed up as follows –

  Null Hypothesis: Every co-efficient is individually equal to zero
  Alternate Hypothesis: Not every co-efficient is equal to zero

$$\mathbf{H_0}: \beta_1 = \beta_2 = \ldots = \beta_k = 0$$
$$\mathbf{H_A}: \beta_1 = \beta_2 = \ldots = \beta_k \neq 0$$

  - Step 1: Create Dummy Variables on Categorical Features to reduce multiple collinearity
  - Step 2: Divide Data into Train and Test
  - Step 3: On Train Data perform Scaling Operations
  - Step 4: Divide Data into X (Independent attributes) and y (target variable)
  - Step 5: Use feature selection method to choose the optimal attributes from X Dataset. Can use a bottom-up approach (choosing single variable, and then adding on), or use RFE (choosing a pool of variables, and eliminating it).
  - Step 6: using the X Dataset above, iteratively build models and compare their metrices – Adjusted R2, R2, F-statistics, P-value, etc.
  - Step 7: Finalize on a model. Check the residuals and other Hypothesis that were considered.
  - Step 8: Considering #7 was successful and hypothesis were verified; on test dataset, perform scaling, divide data into X & y and use the columns identified from the train dataset to create the test model.
  - Step 9: Compute the R2 score for Test dataset. It should be ~ 5% of the train dataset.

2. Explain the Anscombe's quartet in detail.

Ans:

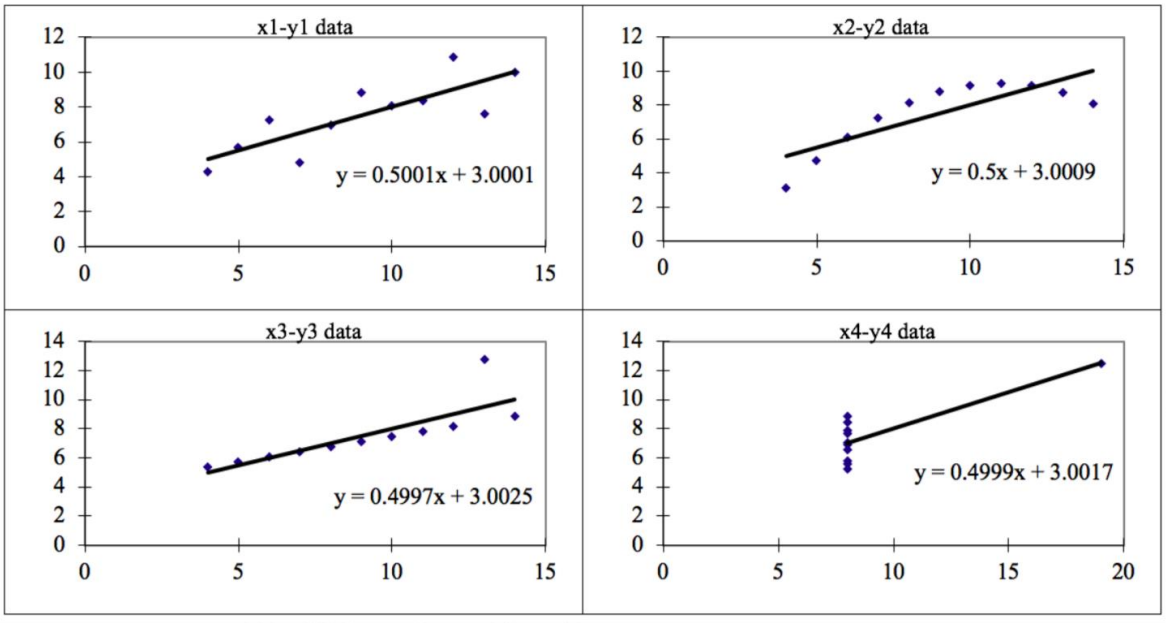
- Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that deceives the regression model if built. They have very different distributions and appear differently when plotted on scatter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

- It illustrates the importance of plotting the graphs before analysing and model building, and the effect of other observations on statistical properties. There are these four data set plots which have nearly same statistical observations, which provides same statistical information that involves variance, and mean of all X, Y points in all four datasets.
- This tells us about the importance of visualising the data before applying various algorithms out there to build models.

- These four plots and their statistical summary can be defined 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	

- When these models are plotted on a scatter plot, all datasets generate a different kind of plot that is not interpretable by any regression algorithm which is misled by these peculiarities and can be seen as follows:



- 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

- This proves the importance of Data Visualization since datasets that have the same statistical summary, might not have the same fit.

3. What is Pearson's R?

Ans:

- 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 correlation coefficient is the covariance of the two variables divided by the product of their standard deviations.
- There are certain requirements for Pearson's Correlation Coefficient:
    - Scale of measurement should be interval or ratio
    - Variables should be approximately normally distributed
    - The association should be linear
    - There should be no outliers in the data

Formula

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

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

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

Ans:

Scaling is –

- A method used to normalize the range of independent variables or features of data
- Also known as data normalization and is generally performed during the data pre-processing step

Scaling is Performed to –

- Handle highly varying magnitudes or values or units that may be present in the data

- It is used to bring all the independent variables into a common standard for the machine learning algorithm to use.

  E.g., A weight of 1000 grams and a price of 10 dollars represents completely two different things for us - but for a model as a feature, it treats both as same.

  In the above case, "Weight" cannot have a meaningful comparison with the "Price." So, the assumption algorithm makes that since "Weight" > "Price," thus "Weight," is more important than "Price."
- In order to reduce such errors, we need to scale the features.

Difference Between Normalized and Standardized Scaling -

Normalized scaling	Standardized scaling
Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
It is used when features are of different scales.	It is used when features are of different scales.
Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
Outliers affect normalized scaling	Standardized scaling is not as affected by outliers as normalized scaling
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.
It is often called as Scaling Normalization	It is often called as Z-Score Normalization.

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

Ans:

- 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) equalling 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).

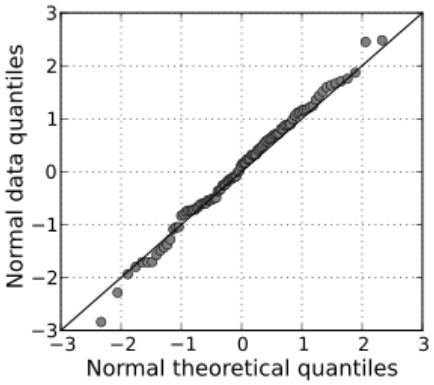
6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Ans:

Q-Q Plots (Quantile-Quantile plots) are:

- Plots of quantiles of two datasets against each other
- 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 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.
- In linear regression, Q-Q plots are used to visually check that the data meets the normality assumptions of linear regression.

Usage of Q-Q Plots -

- Q–Q plots are used as a graphical means of estimating parameters in a location-scale family of distributions.
- Q-Q plots let us check that the data meet the assumption of normality. They compare the distribution of your data to a normal distribution by plotting the quartiles of our data against the quartiles of a normal distribution. If our data are normally distributed, then they