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?
2. Why is it important to use drop_first=True during dummy variable creation?
3. Looking at the pair-plot among the numerical variables, which one 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?
5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

General Subjective Questions

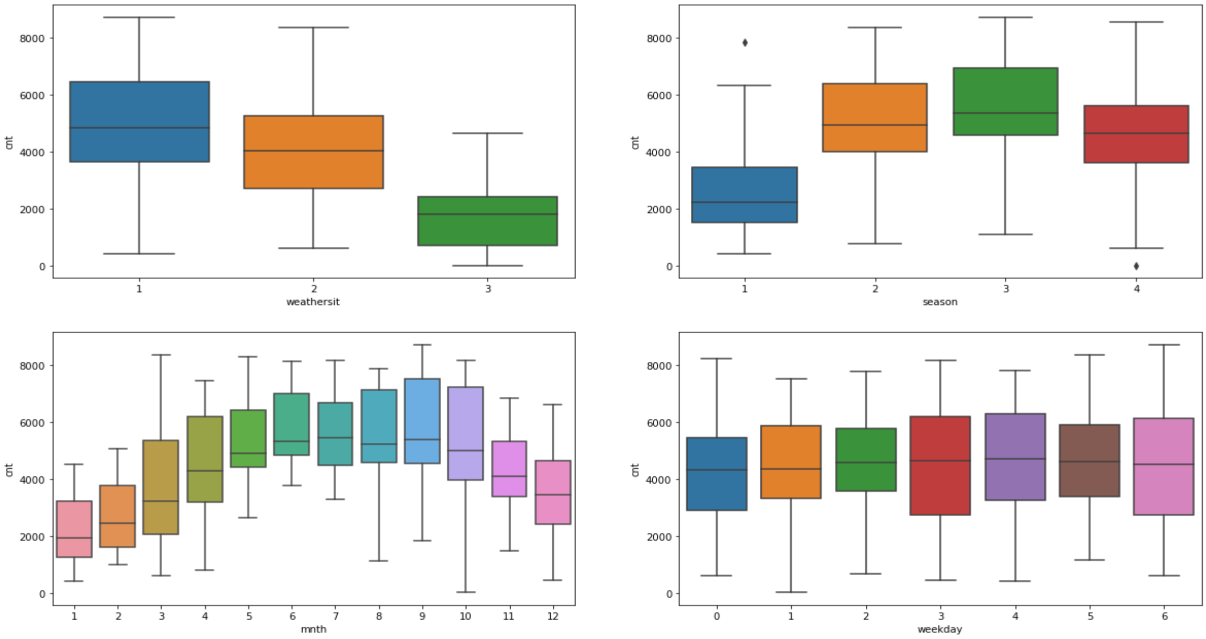
1. Explain the linear regression algorithm in detail.
2. Explain the Anscombe's quartet in detail.
3. What is Pearson's R?
4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

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?

Answer:

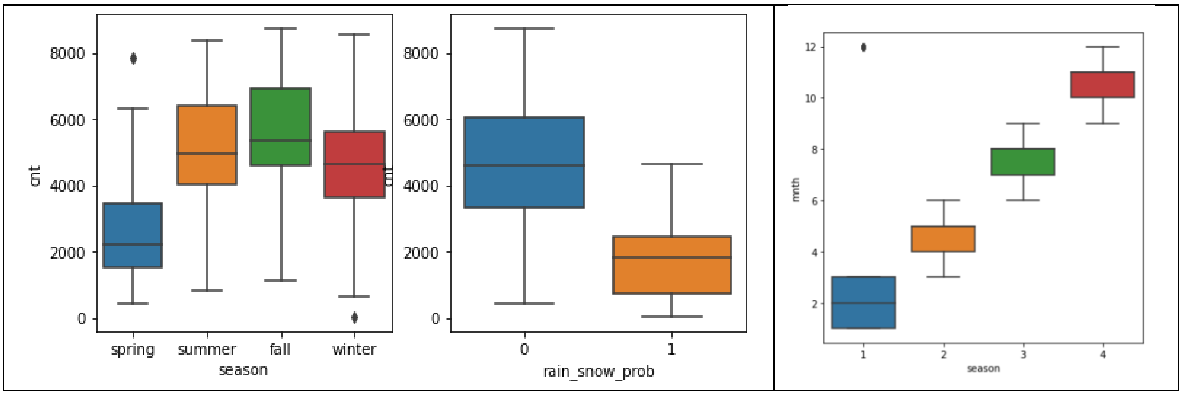
Univariate Analysis of the categorical variables:



The categorical variables considered for the analysis were `weathersit`, `season`, `mnth` & `weekday`.

Bivariate Analysis on the derived categorical variables:

`rain_snow_prob` is a derived variable from `weathersit` variable as per the data dictionary.



Inference:

- When the likely hood of rain or snow is less, more bikes were hired (the target/dependent variable is relatively much higher)
- During spring season less bikes were hired (the target/dependent variable is relatively much lower)
- 3 months of the year, the bikes hired were low and increases in the next 6 months. 75% percentile of the target variable `cnt` is the highest in month 9.

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

Answer:

drop_first=True drops the first column during dummy variable creation. It helps by reducing the extra column created during dummy variable creation. Thereby it reduces the correlations created among dummy variables.

In the assignment, creating the dummies for `season_type` created 2 columns. Using drop_first=True one column was dropped.

```
# Get the dummy variables for the derived feature 'season_type' and store it in a new variable - 'Sea_Stat'
Sea_Stat = pd.get_dummies(bikes.season_type)
#Print and check the dummiesprint (Sea_Stat.value_counts())
```

- `10` normal_season
- `01` off_season

```
# Dropping the first column `normal_season`
Sea_Stat = pd.get_dummies(bikes.season_type, drop_first = True)
#Print and check the dummies
print (Sea_Stat.value_counts())
```

```
off_season
0             545
1             171
dtype: int64
```

- `10` normal_season is now `0`
- `01` off_season is now `1`

---

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

Answer:

After cleaning the data: Looking at the pair-plot among the numerical variables in this assignment, the variable `atemp` has the highest correlation with the target variable `cnt`.

Top 2 numerical variables that has high correlation with the target variable

Numerical Variable	Correlation score with target variable \`cnt\`
\`atemp\`	0.66
\`yr\`	0.59

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

Answer:

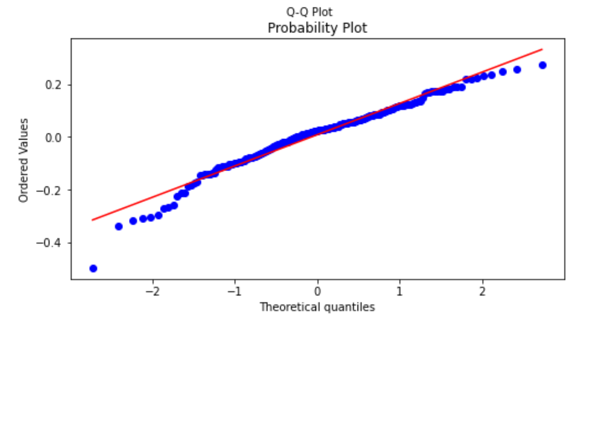
Assumptions of Linear Regression that needs to be checked after building the model:

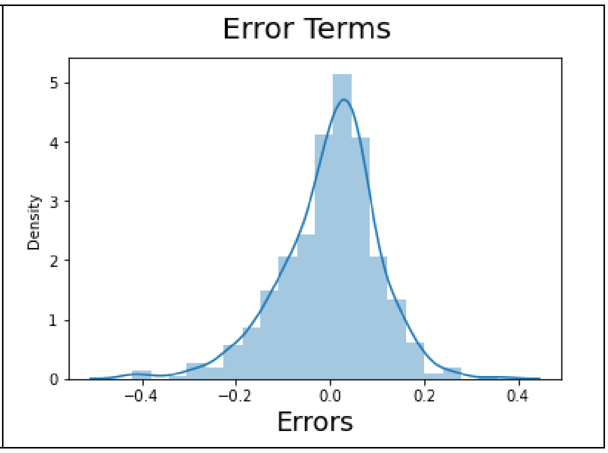
- Homoscedasticity
- Mean of the Error centres near Zero
- Error distribution is normal

Validation of assumptions after building the model on the training set:

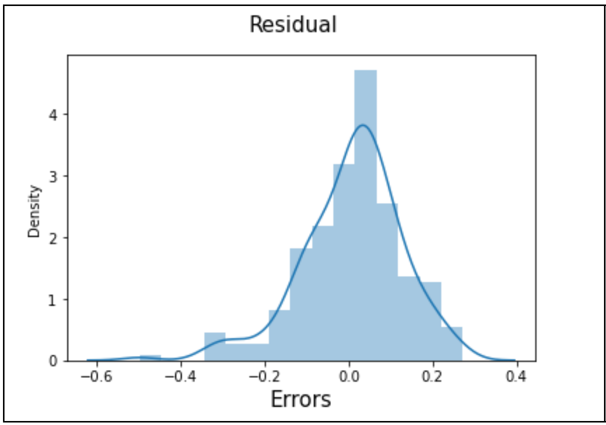
Q-Q Plot to visually check and confirm - Homoscedasticity and Normality

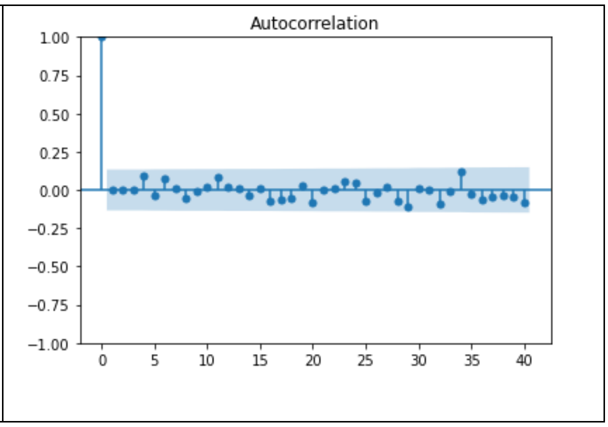
Histogram Plot of Error Terms to understand the distribution





After Prediction: Residual analysis





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

Answer:

OLS Regression Results:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    cnt   R-squared:                       0.791
Model:                            OLS   Adj. R-squared:                  0.789
Method:                 Least Squares   F-statistic:                     468.7
Date:                Wed, 14 Sep 2022   Prob (F-statistic):          6.44e-167
Time:                        09:40:53   Log-Likelihood:                 426.01
No. Observations:                 501   AIC:                            -842.0
Df Residuals:                     496   BIC:                            -820.9
Df Model:                           4
Covariance Type:            nonrobust
==================================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------
const              0.2197      0.018     12.360      0.000       0.185       0.255
yr                 0.2408      0.009     25.730      0.000       0.222       0.259
atemp              0.4246      0.028     15.412      0.000       0.370       0.479
rain_snow_prob    -0.2457      0.031     -8.008      0.000      -0.306      -0.185
off_season        -0.1592      0.014    -11.375      0.000      -0.187      -0.132
==============================================================================
Omnibus:                       49.195   Durbin-Watson:                   2.146
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               83.424
Skew:                          -0.634   Prob(JB):                     7.67e-19
Kurtosis:                       4.546   Cond. No.                         9.56
==============================================================================
```

Variance Inflation Factor – VIF:

	Features	VIF
0	yr	1.98
1	atemp	1.90
3	off_season	1.12
2	rain_snow_prob	1.02

Based on the above, the below 3 features contributing significantly towards explaining the demand of the shared bikes:

1. **yr**
   - from 2018, the demand has increased in 2019
2. **atemp**
   - feeling temperature that potentially can include humidity and windspeed, instead of only the temp
3. **rain_snow_prob** // This is a derived variable from independent variable **weathersit**
   - when there is a possibility for rain/snow, the demand of shared bikes goes down

General Subjective Questions-Answers

1. Explain the linear regression algorithm in detail.

Answer:

Linear Regression algorithm is a simplest form of “Regression” based machine learning algorithm. In regression-based algorithm the target/output/dependent variable is a continuous variable.

The linear regression algorithm is mostly used for finding the relationship between the “Target/Dependent” variable and “Predictor/Independent” variables. And is also used in forecasting the target variable based on the predictor/independent variables.

In this case, the target/dependent and the predictor/independent variable(s) are linearly corelated.

Examples:

- Predicting the sales based on previous sales data
- Predicting the number of customers in a Shopping mall based on the past data
- Predicting the commodity price based on the past commodity data

Types of Linear Regression:

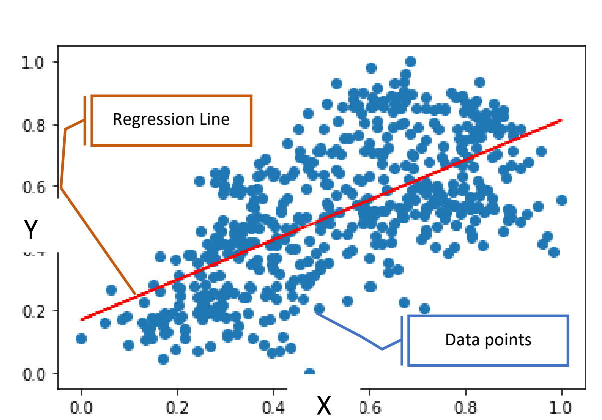
- SLR: Single Linear Regression
    - It is a statistical technique that uses one predictor/independent variable to find relationship & predict the outcome of a target variable
- MLR: Multiple Linear Regression – it’s an extension of SLR
    - It is a statistical technique that uses several predictor/independent variables to find relationship & predict the outcome of a target variable

Mathematical Formula & Visualization:

SLR: $Y = \beta_0 + \beta_1 X$

MLR: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + E$

Y is the Target/Output/Dependent variable
β_0 is the Intercept
β_1, β_2, β_n are the slopes of X_1, X_2, X_n respectively
X_1, X_2, X_n are the predictor/independent variables
In case of SLR, only one independent variable X
E – Error terms



2. Explain the Anscombe's quartet in detail.

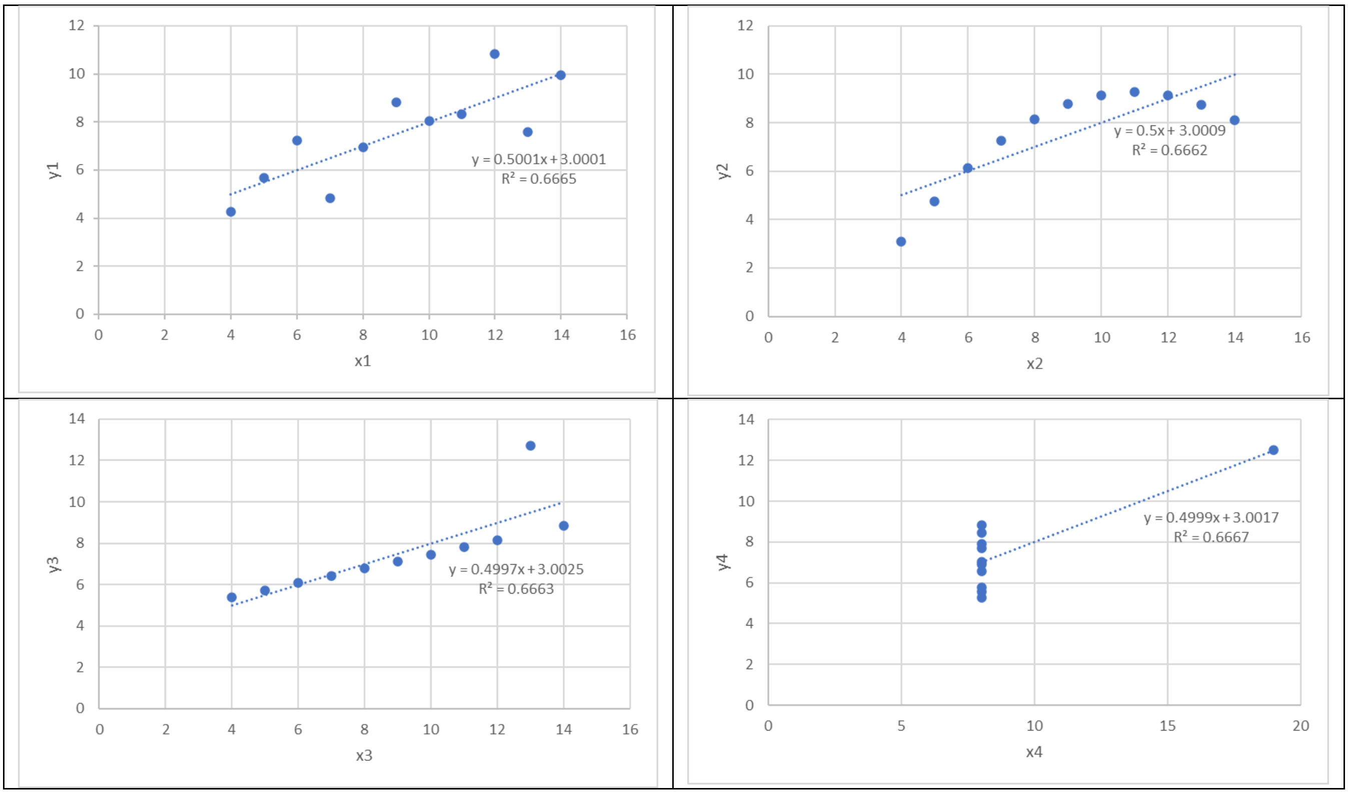
Answer:

Anscombe's quartet is dataset with shape (11, 8) with 4 x and 4 y.

Anscombe dataset:

	x1	x2	x3	x4	y1	y2	y3	y4
0	10	10	10	8	8.04	9.14	7.46	6.58
1	8	8	8	8	6.95	8.14	6.77	5.76
2	13	13	13	8	7.58	8.74	12.74	7.71
3	9	9	9	8	8.81	8.77	7.11	8.84
4	11	11	11	8	8.33	9.26	7.81	8.47
5	14	14	14	8	9.96	8.10	8.84	7.04
6	6	6	6	8	7.24	6.13	6.08	5.25
7	4	4	4	19	4.26	3.10	5.39	12.50
8	12	12	12	8	10.84	9.13	8.15	5.56
9	7	7	7	8	4.82	7.26	6.42	7.91
10	5	5	5	8	5.68	4.74	5.73	6.89

Scatter Plot of the dataset:



Insights:

x1, y1: Linear regression fits	**x2, y2:** this couldn't be fitted with the linear line and looks polynomial
x3, y3: Outlier doesn't fit the Linear line	**x4, y4:** Linear doesn't fit at all

Though the R2 value and the equation are almost same ***[y = 0.5x + 3.00 and R^2 = 0.67]***, the Linear model doesn't fit 3 sets and cannot be used to interpret the relationship and cannot be used to predict.

Inference from the Anscombe's quartet is that though the numerical values are same, it is significant to plot them as graphs to visualize before analysing and making the model.

---

3. What is Pearson's R?

Answer:

Pearson's R is also called as Pearson correlation coefficient. The Pearson's R measures the linear association between random x and y variables. It is also used to eliminate the predictor/independent variables that are highly correlated to each other.

Mathematical Formula:

Pearson's R equals to covariance of x, y divided by the product of their standard deviations.

Pearson's R = cov(x,y)/ (σx* σy)

σ – standard deviation

cov - covariance

numpy corrcoef() method or

scipy pearsonr() or

pandas dataframe.corr() can be used to find the pairwise correlation

Code snippet

```
from scipy.stats import pearsonr
```

```
# Apply the pearsonr() on anscombe's dataset
corr, _ = pearsonr(x1, y1)
print('Pearsons correlation on anscombe\'s x1, y1 : %.3f' % corr)
```

```
Pearsons correlation on anscombe's x1, y1 : 0.816
```

Interpretation: The value ranges from -1 to +1.

Pearson's R value	Interpretation
-1	Means highly negatively correlated
0	Means No linear correlation
1	Means highly positively correlated

If the sample has more noise the score will be positively or negatively less.

Note:

- The Pearson's R is used to identify patterns between x, y.
- But the R^2 score is used to identify the strength of a model.

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

Answer:

Scaling is a method to normalize or bring the x, y variables/features closer to a comparable scale or a fixed range. It is one of the important data pre-processing steps before modelling.

Many of the algorithms are highly sensitive to the scale of the variables and the algorithm can be biased towards higher scale variables.

There are two types of scaling: Normalized and Standardized Scaling

Normalized Scaling	Standardized Scaling
Also called as min-max scaling	Also called as z-score normalization
Min and Maximum value of the variables used for scaling	Mean and Standard deviation used for scaling
Scale range: 0 to 1 or -1 to +1	Mean of 0 and standard deviation of 1
Outliers affect the scaling as it is based on Min,Max	Outliers don't affect this scaling as it based on the mean and standard deviation
Used when the ML algorithms such as support vector machines (SVM) and k-nearest neighbours (KNN) where distance between the data points is important	Used when the ML algorithm assumes that data is normally distributed
from sklearn.preprocessing import MinMaxScaler	from sklearn.preprocessing import StandardScaler

---

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

Answer:

Formula of VIF is

VIF = 1 / (1- R^2)

If the R^2 is equal to 0, VIF will be 1.

If the R^2 is equal to 1, VIF will be infinity

VIF = 1 / (1-1)
= 1 / 0
= infinity

R^2 will be 1 when there is a perfect correlation between two independent variables.

In this case, one of the variables has to be dropped from the dataset which is the cause of the perfect multicollinearity.

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

Answer:

Q-Q Plot is a probability plot. It is a quantile-quantile plot comparing the quantiles against each other. Normal distribution of the residuals can be validated by the Q-Q plot.

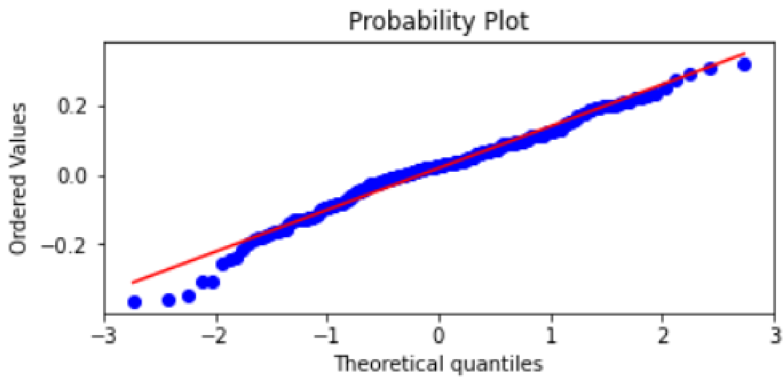
Q-Q plot helps in determining if a dataset follows normal, uniform, exponential probability distribution, etc.

Usage 1 in Linear Regression:

Assumption of Linear Regression:

- **Homoscedasticity** of Residual and
- **Normal distribution** of the Error terms

The Q-Q plot on the residual [difference between the test data & predicted data] will visually confirm if the above assumptions of linear regression are met. Approximately a straight line [45 degree] represents that the data is normally distributed.



Usage 2 in Linear Regression:

In case of Linear Regression, if the training data set and the test data set are separately received, then the Q-Q plot will help in confirming that both data sets are from populations with same distribution.