Linear Regression

Subjective Questions & Answers on 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?

Please take a look at the following correlation table between dummy variables of weather and season against the target variable cnt.

Season	**Correlation w/ cnt**	**Weather**	**Correlation w/ cnt**
Summer	0.111492	Cloudy	-0.143372
Spring	-0.545842	Rainy	-0.220788
Winter	0.040317		

From the above correlation table, we can see that spring has the strongest negative correlation with value of -0.545842, we can say that during spring people has the tendencies of not taking a bike sharing.

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

We remove the first dummy variable because it is already represented with the rest of the dummy variables, and therefore we are reducing correlations between the variables, consider an example from housing data:

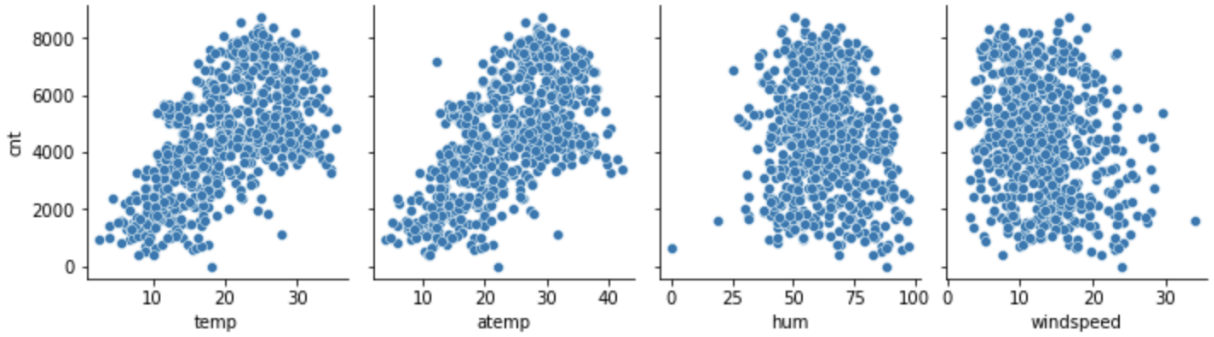
	Unfurnished	Semi-Furnished	Furnished
Unfurnished	1	0	0
Semi-Furnished	0	1	0
Furnished	0	0	1

Given we have semi-furnished and furnished as dummy variable, we no longer need unfurnished, because its already represented by semi-furnished and furnished, that is if semi-furnished and furnished is 0, it means that the house is unfurnished.

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

Following are the pair plot of all numerical variables against the target variable, there are only a handful of numerical variables, as the other variables with seemingly numeric values are actually categorical variables.

As for casual and registered, we are not considering them as part of the independent variables because they are actually part of the target variable, that is `cnt` is the total of `casual` and `registered` values.



From the above plot, we can see that there is a correlation between `temp` and `atemp` with `cnt`, but it is unclear which one is the highest, to see the highest correlation, we would need to calculate them.

	temp	atemp	hum	windspeed
cnt	0.627044	0.630685	0.098543	0.235132

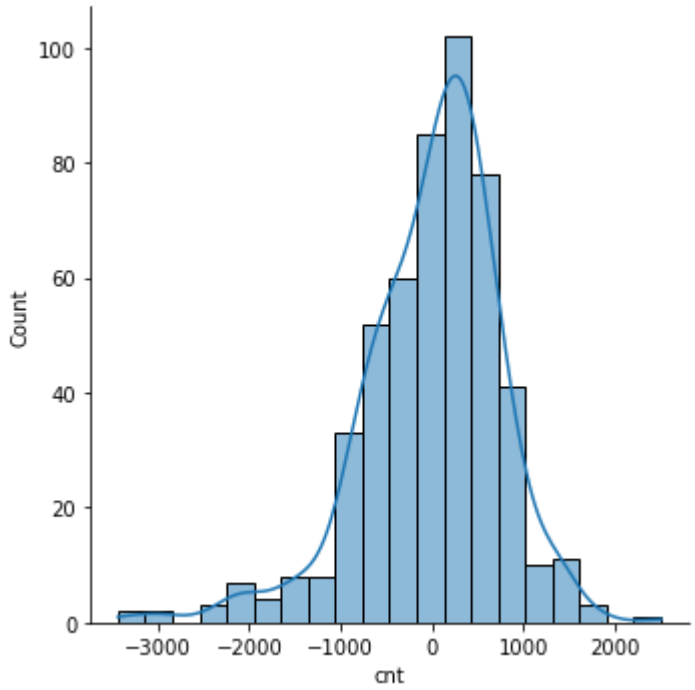
Now we can see clearly that the highest correlation is between `cnt` and `atemp` with 0.630685, while `cnt` and `temp` is just slightly below them.

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

We validate our assumption of our linear regression model by running residual analysis, that is plot the error terms of our model, and look if it form standard distribution, that is mostly centered around 0.

We can plot the distribution using the following statements as an example:

```
y_train_pred = model.predict(X_train)
sns.displot((y_train.cnt - y_train_pred), bins = 20, kde=True)
```



From the above chart, it is clear that our assumption of the model has been validated because the error terms are distributed around the center, that is mostly close to zero and form standard distribution.

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

Lets take a look at the following screenshot of coefficients from our notebook:

Coeficients

Lets output the coeficients of our model.

```
dict(zip(X_train.columns[model.support_], model.estimator_.coef_[0]))
```

```
{'yr': 1902.4137854971468,
 'temp': 6125.121529066783,
 'hum': -1663.7146901903857,
 'windspeed': -1469.6705714825362,
 'spring': -1050.6253260417755,
 'rainy': -1337.132022150946,
 'apparent_temperature': -1887.166531996481,
 'windchill': -698.8550618924029}
```

Including derived metrics that we have generated, there are temp, yr and apparent_temperature that is the most contributing variables. But, excluding our derived metrics, the top 3 contributing variables would be: temp, yr and hum.

General Subjective Questions

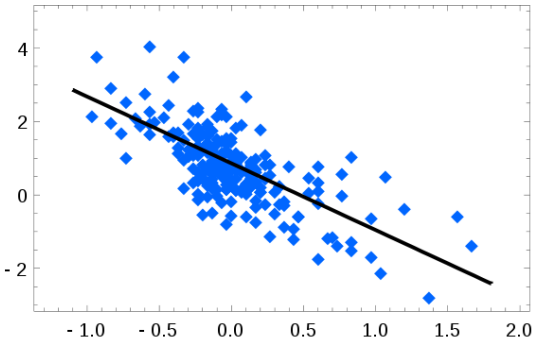
1. Explain the linear regression algorithm in detail.

Linear regression is a way of modelling relationship between dependent and independent variables, in the case of a single independent variable, we called it simple linear regression, for more than one independent variables, we called it multiple linear regression.

The formula of the algorithm is as follows:

$$y_i = \beta_0 + \beta_1 x_{i1} + \ \ldots \ + \beta_p x_{ip} + \varepsilon_i$$

Linear model can be fit in many ways, one of them is minimizing the square error of the residual, it is also called ordinary least squares, once fitted, the model may form a line as such.



The model can be fit fairly easily in python, following is simplification of how fitting a linear regression in python using scikit-learn, some operations are deliberately omitted to just focus on the model fitting and evaluation.

```
from sklearn.linear_model import LinearRegression

model = LinearRegression()
model.fit(X_train, y_train)
print("R-square train: ", model.score)
```

To evaluate the model, we can read the R^2 of the test data as follows:

```
from sklearn.metrics import r2_score

y_pred = model.predict(X_test)
print("R-square test: ", r2_score(y_train, y_pred))
```

If the R-square of train data is way less than the test data, it is called underfit. If R-square the train data is way more than the test data, it is called overfit. We have to bring the gap closer so that between train and test data have a similar R-square.

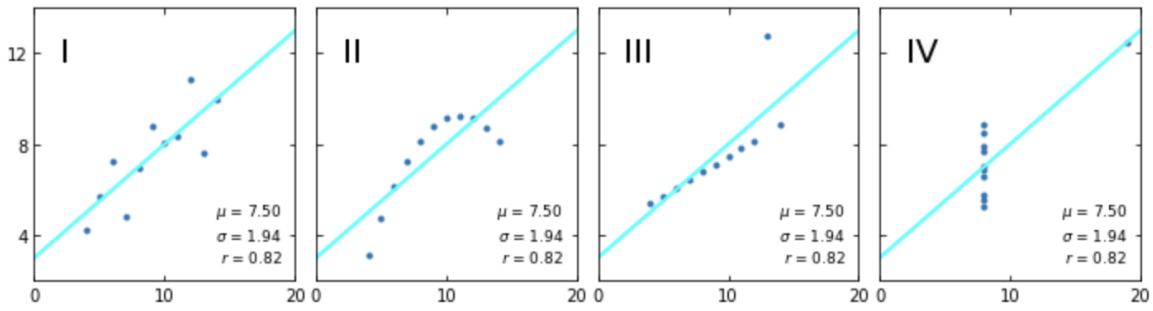
2. Explain the Anscombe's quartet in detail.

Anscombe's Quartet is a collection of four data sets that have almost identical simple descriptive statistics, but have very different distributions and they show up very differently when plot in a chart.

Following are the datasets, with x is the same for the first three datasets.

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

Following is the visual of the data in a chart.



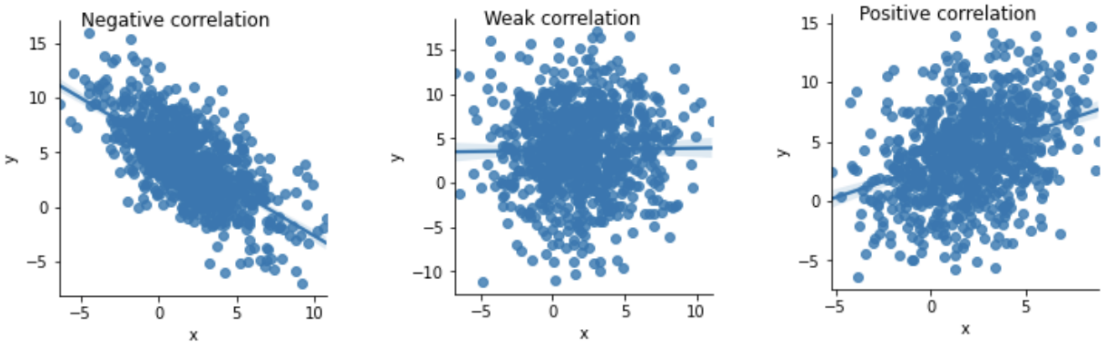
Despite the differences in the distributions and visual representation in the chart, as you can see, their descriptive statistics, which are mean, standard deviation and correlation are the same.

These datasets were constructed by Francies Anscombe to emphasize the importance of visualization when analyzing the data, and the impact of outliers on statistical properties.

3. What is Pearson's R?

Person's R is also known as correlation coefficient, it explains the linear correlation between two datasets, with range of value between -1 to 1.

Following is a visual representation of how correlation in different datasets.



The formula can be written as follows:

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

It may look first hard to grasp, but it simply says that they are the sums of the product of z score of x and y, divided by n - 1, we don't have to calculate them by hand because computer programs can do so with speed and accuracy.

In pandas we can calculate the above dataset as follows:

```
df.corr()
```

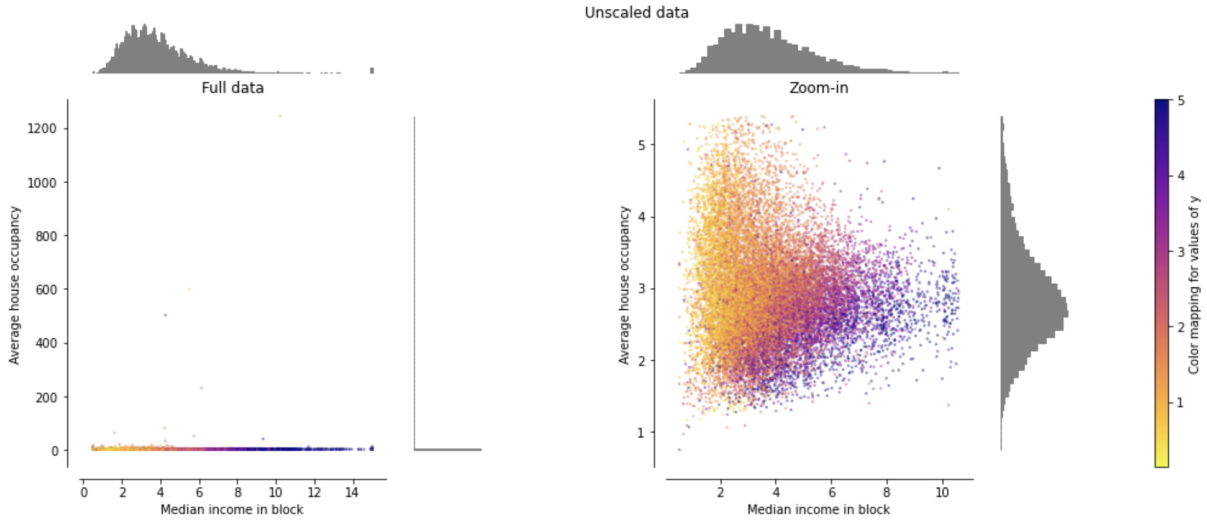
And it will display their correlation as such, from left to right, negatively correlated, weak correlation and positive correlation.

```
          x         y
x  1.000000 -0.665475
y -0.665475  1.000000

          x         y
x  1.000000  0.015038
y  0.015038  1.000000

          x         y
x  1.000000  0.357675
y  0.357675  1.000000
```

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

Scaling is way to transform our data so that they are distributed proportionally between each features. It needs to be performed so that each feature contribution is proportional, it can also causes gradient descent to converge much faster, its also important if regularization is used as the loss function so that each coefficient are penalized accordingly, and finally it provides an interpretability on the coefficients.

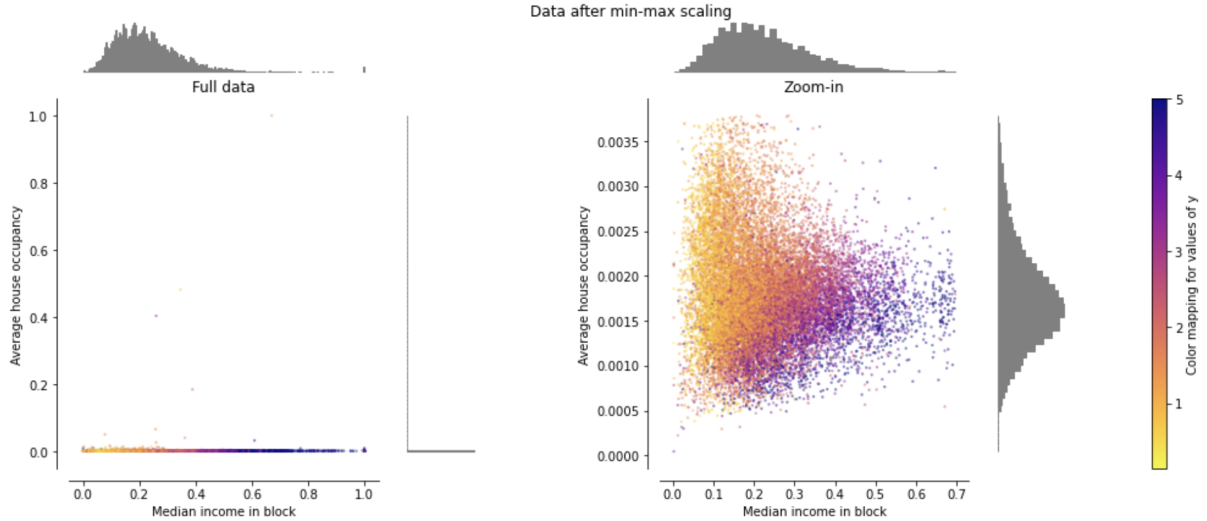
Lets take an example with the following dataset, on the left side is the full dataset, on the right is the dataset zoomed in to exclude outliers.



Normalized scaling would mean reducing range of the value down to just between 0 and 1, it is also known as MinMaxScaler, you can get their values with the following formula:

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

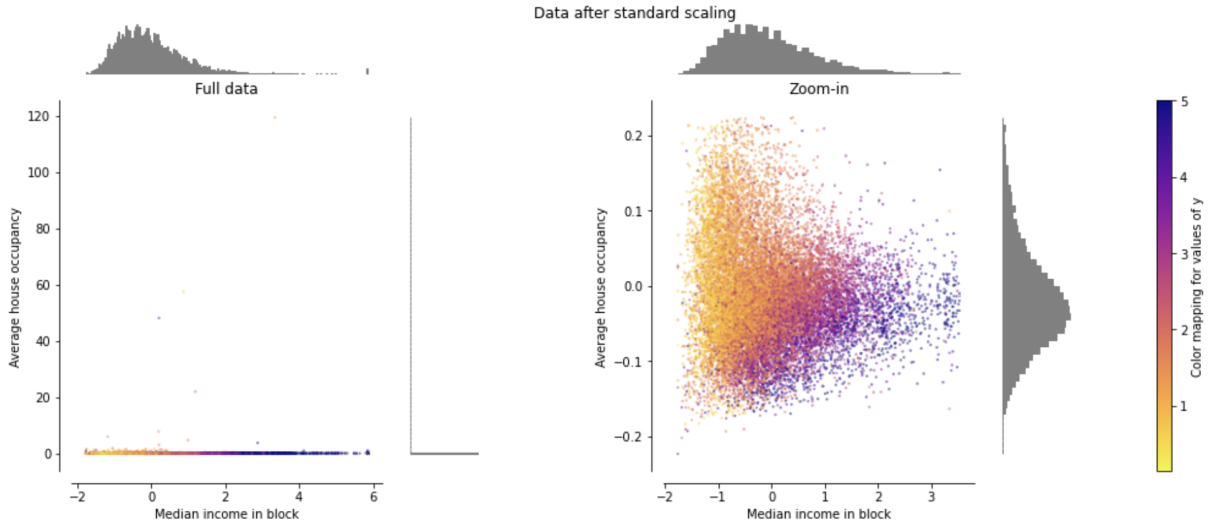
If we were to normalize with min max scaler, it would appear as follows, as you can see, on the left the data is squeezed to just between 0 and 1, and on the right, we can see the majority of the data sits between 0 and 0.005.



Standardized scaling on the other hand would transform the data have zero mean and unit-variance, you can get their values with the following formula:

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

If we were to normalize with the standard scaler, it would appear as follows, as you can see, on the zoomed in chart on the right it shows the data have zero mean, and on the left it shows that it is distributed with unit-variance.



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

Before we get down to the answer, lets have a look at the formula:

$$VIF_i = \frac{1}{1-R^2}$$

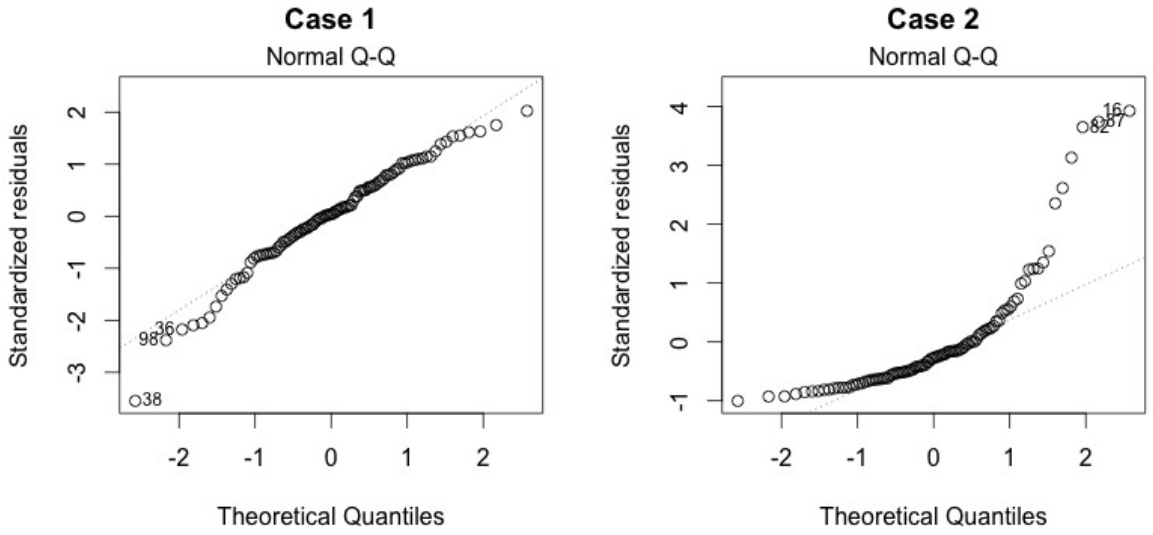
Lets experiment with a correlation of 1, that is positively perfect calculation, VIF would be $\frac{1}{0}$, which is undefined. But if we calculate the limit of the $\frac{1}{x}$as x approaches zero from left or right is positive and negative infinity.

VIF provides an index to measure the severity of multicollinearity of a particular feature, and therefore as VIF approaches infinity when correlation approaches positive or negative 1, it explains the severity index of the multicollinearity of the feature.

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

Q-Q plot is two sets of quantiles plot in a scatterplot against each other, if both dataset came from the same distribution, we should see roughly a straight line formed by the points.

Q-Q plot can be used as one of the tool to run diagnostic of a given model, we can plot the residuals and see if they are normally distributed, do they form a straight line or diverge significantly.



From the above example, we can see that the first case would seem just fine because they follow normal distribution, but on the second case the points diverge significantly and there may be a potential problem on the model.