Module 6 Exercises - Correlation and Models

Exercise 1:

From the datasets folder, load the "tamiami.csv" file as a dataframe. Rename the columns (in order) to the following:

- location
- sales
- employees
- restaurants
- foodcarts
- price

Then do a correlation table on that dataframe. What features (columns) are correlated? What features aren't correlated?

In [1]:
```
import pandas as pd
import numpy as np
```

In [4]:
```
Location = "datasets/tamiami.csv"
df = pd.read_csv(Location)

df.head()
```

Out[4]:

	Cart Location	Hot Dog Sales	Employees in Nearby Office Buildings	Num of Nearby Restaurants	Num of Other Food Carts Nearby	Price
0	1	100	1600	8	12	4.16
1	2	80	1200	6	13	4.63
2	3	450	2800	19	6	0.50
3	4	580	4300	19	2	0.47
4	5	100	1400	6	13	4.24

In [14]:
```
df.rename(columns={'Cart Location':'location', 'Hot Dog Sales':'sales',
                   'Employees in Nearby Office Buildings':'employees',
                   'Num of Nearby Restaurants': 'restaurants',
                   'Num of Other Food Carts Nearby':'foodcarts'},
                   inplace=True)
df.head()
```

Out[14]:

	location	sales	employees	restaurants	foodcarts	Price
0	1	100	1600	8	12	4.16
1	2	80	1200	6	13	4.63
2	3	450	2800	19	6	0.50
3	4	580	4300	19	2	0.47
4	5	100	1400	6	13	4.24

Correlation

A correlation coefficient is a numerical measure of some type of correlation. A correlation is a statistical relationship between two variables. The variables may be two columns of a given data set of observations, often called a sample, or two components of a multivariate random variable with a known distribution.

```
df.corr()
```

Compute pairwise correlation of columns, excluding NA/null values.

As tools of analysis, correlation coefficients present certain problems, including the propensity of some types to be distorted by outliers and the possibility of incorrectly being used to infer a causal relationship between the variables

The linear correlation coefficient measures the strength and direction of a linear relationship between two variables. The linear correlation coefficient is sometimes referred to as the Pearson product moment correlation coefficient

The value of a correlation coefficient, r, is such that -1 < r < +1.
The + and – signs are used for positive linear correlations and negative linear correlations, respectively.

- Positive correlation If x and y have a strong positive linear correlation, r is close to +1. An r value of exactly +1 indicates a perfect positive fit. Positive values indicate a relationship between x and y variables such that as values for x increases, values for y also increase.
- Negative correlation If x and y have a strong negative linear correlation, r is close to -1. An r value of exactly -1 indicates a perfect negative fit.
  Negative values indicate a relationship between x and y such that as values for x increase, values for y decrease.
- No correlation: If there is no linear correlation or a weak linear correlation, r is close to 0. A value near zero means that there is a random, nonlinear relationship between the two variables

A perfect correlation of ± 1 occurs only when the data points all lie exactly on a straight line. If r = +1, the slope of this line is positive. If r = -1, the slope of this line is negative.

Strong and Week Correlation

A correlation greater than 0.8 is generally described as strong, whereas a correlation less than 0.5 is generally described as weak.

These values can vary based upon the "type" of data being examined. A study utilizing scientific data may require a stronger correlation than a study using social science data.

In [17]:
```
# Create a table of correlation values
# No correlation exists when the value = 0
# A direct relationship exists when the value = 1
df.corr()
```

Out[17]:

	location	sales	employees	restaurants	foodcarts	Price
location	1.000000	0.042705	-0.068923	0.049701	0.077219	-0.138444
sales	0.042705	1.000000	0.943238	0.913674	-0.919762	-0.966378
employees	-0.068923	0.943238	1.000000	0.856976	-0.874692	-0.881540
restaurants	0.049701	0.913674	0.856976	1.000000	-0.761793	-0.933951
foodcarts	0.077219	-0.919762	-0.874692	-0.761793	1.000000	0.860154
Price	-0.138444	-0.966378	-0.881540	-0.933951	0.860154	1.000000

What features (columns) are correlated? What features aren't correlated?

In [7]:
```
#All of them seem to have a strong positive and negative coorination. if you have
```

Exercise 2:

Using the dataframe from the previous exercise, choose features (columns) to create a linear regression formula to predict sales. Try it with and without the y-intercept. How does it make a difference? Does adding or removing features in your model formula make a difference in the output?

In [8]:
```
#use this library to build a statistical test for linear regression
import statsmodels.formula.api as smf
```

In [9]:
```
result = smf.ols('sales ~ location + Price + foodcarts', data=df).fit()
```

In [10]:
```
result.summary()
```

Out[10]:

OLS Regression Results

Dep. Variable:	sales	**R-squared:**	0.965
Model:	OLS	**Adj. R-squared:**	0.961
Method:	Least Squares	**F-statistic:**	237.0
Date:	Wed, 21 Aug 2019	**Prob (F-statistic):**	5.39e-19
Time:	10:42:50	**Log-Likelihood:**	-147.02
No. Observations:	30	**AIC:**	302.0
Df Residuals:	26	**BIC:**	307.7
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	581.1520	18.859	30.816	0.000	542.388	619.916
location	-0.5833	0.808	-0.722	0.477	-2.243	1.077
Price	-75.7789	8.584	-8.828	0.000	-93.423	-58.135
foodcarts	-12.7023	3.128	-4.061	0.000	-19.132	-6.273

Omnibus:	0.003	**Durbin-Watson:**	2.146
Prob(Omnibus):	0.999	**Jarque-Bera (JB):**	0.158
Skew:	-0.015	**Prob(JB):**	0.924
Kurtosis:	2.645	**Cond. No.**	59.7

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

In [11]:
```
result = smf.ols('sales ~ location + Price + foodcarts-1', data=df).fit()
```

In [12]:
```
result.summary()
```

Out[12]:

OLS Regression Results

Dep. Variable:	sales	**R-squared:**	0.593
Model:	OLS	**Adj. R-squared:**	0.548
Method:	Least Squares	**F-statistic:**	13.13
Date:	Wed, 21 Aug 2019	**Prob (F-statistic):**	1.78e-05
Time:	10:42:54	**Log-Likelihood:**	-201.40
No. Observations:	30	**AIC:**	408.8
Df Residuals:	27	**BIC:**	413.0
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
location	11.5117	4.242	2.714	0.011	2.807	20.216
Price	-86.1379	51.559	-1.671	0.106	-191.928	19.652
foodcarts	23.3037	17.441	1.336	0.193	-12.483	59.090

Omnibus:	3.998	**Durbin-Watson:**	0.795
Prob(Omnibus):	0.135	**Jarque-Bera (JB):**	3.416
Skew:	0.820	**Prob(JB):**	0.181
Kurtosis:	2.785	**Cond. No.**	28.4

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

In [ ]:
```
```