Aim :- Implementation of Statistical Hypothesis Test using Scipy and Sci-kit learn

Todo :-

Correlation Tests:
1. Pearson's Correlation Coefficient
2. Spearman's Rank Correlation
3. Kendall's Rank Correlation:
4. Chi-Squared Test

About Dataset:

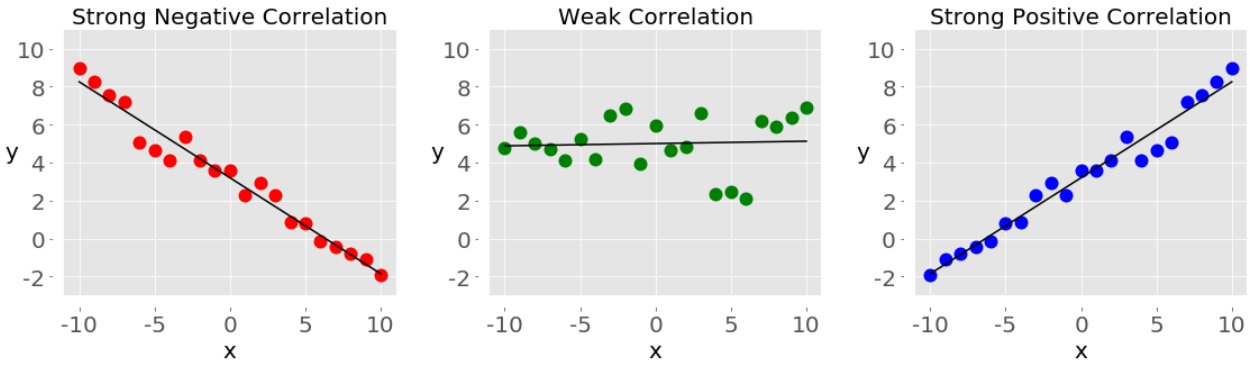
The data set contains customer demographics (age, gender, marital status, citytype, stay in current city), product details (productid and product category) and total purchase amount from last month. From this dataset a company can understand the customer purchase behaviour (specifically, purchase amount) against various products of different categories. From this dataset we can build a model to predict the purchase amount of customer against various products which will help them to create personalized offer for customers against different products.

Theory :-

Correlation

Statistics and data science are often concerned about the relationships between two or more variables (or features) of a dataset. Each data point in the dataset is an observation, and the features are the properties or attributes of those observations.

If you analyze any two features of a dataset, then you'll find some type of correlation between those two features. Consider the following figures:



Each of these plots shows one of three different forms of correlation:

1. **Negative correlation :** In the plot on the left, the y values tend to decrease as the x values increase. This shows strong negative correlation, which occurs when large values of one feature correspond to small values of the other, and vice versa.

2. **Weak or no correlation :** The plot in the middle shows no obvious trend. This is a form of weak correlation, which occurs when an association between two features is not obvious or is hardly observable.
3. **Positive correlation:** In the plot on the right, the y values tend to increase as the x values increase. This illustrates strong positive correlation, which occurs when large values of one feature correspond to large values of the other, and vice versa.

Pearson's Correlation Coefficient:

In statistics, the Pearson correlation coefficient — also known as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), the bivariate correlation, or colloquially simply as the 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 normalized 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 relationships or correlations. As a simple example, one would expect the age and height of a sample of teenagers from a high school to have a Pearson correlation coefficient significantly greater than 0, but less than 1. It is a measure of the linear relationship between two random variables - X and Y. Mathematically, if (σXY) is the covariance between X and Y, and (σX) is the standard deviation of X, then the Pearson's correlation coefficient ρ is given by:

$$\rho_{X,Y} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

Spearman's Rank Correlation

In statistics, Spearman's rank correlation coefficient or Spearman's ρ, named after Charles Spearman is a nonparametric measure of rank correlation (statistical dependence between the rankings of two variables). It assesses how well the relationship between two variables can be described using a monotonic function.

A monotonic relationship is a relationship that does one of the following:

(1) as the value of one variable increases, so does the value of the other variable, OR,

(2) as the value of one variable increases, the other variable value decreases. But, not exactly at a constant rate whereas in a linear relationship the rate of increase/decrease is constant.

While Pearson's correlation coefficient is used to measure the relationship between pairs of variables with a linear relationship, Spearman's rank correlation coefficient assesses monotonic relationships, whether they're linear or not. It's a non-parametric measure of correlation, meaning that it doesn't assume that the data is normally distributed.

Kendall's Rank Correlation

Also commonly known as "Kendall's tau coefficient". Kendall's Tau coefficient and Spearman's rank correlation coefficient assess statistical associations based on the ranks of the data. Kendall rank correlation (non-parametric) is an alternative to Pearson's correlation (parametric) when the data you're working with has failed one or more assumptions of the test. This is also the best alternative to Spearman correlation (non-parametric) when your sample size is small and has many tied ranks. Kendall rank correlation is used to test the similarities in the ordering of data when it is ranked by quantities. Other types of correlation coefficients use the observations as the basis of the correlation, Kendall's correlation coefficient uses pairs of observations and determines the strength of association based on the patter on concordance and discordance between the pairs.

Chi-Squared Test

The Chi-square test is a statistical test used to determine the relationship between the categorical variables/columns in the dataset. It examines the correlation between the variables which do not contain the continuous data.

Chi-squared test is a hypothesis test that is used to determine whether there is a significant association between two categorical variables in the data. The test involves two hypotheses (H0 & H1):

- H0 : The two categorical variables have no relationship (independent)
- H1 : There is a relationship (dependent) between two categorical variables

So as a null hypothesis, we keep the positive aspect of the test and in the alternate hypothesis, we keep the negative aspect. The positive aspect of chi-square is that there should not be any correlation because correlation can result in overfitting of the machine learning algorithm. The negative is that there is a correlation between the two categorical columns.

Code:

Pearson's Correlation

```
pearsoncorr = df.corr(method='pearson')
pearsoncorr
```

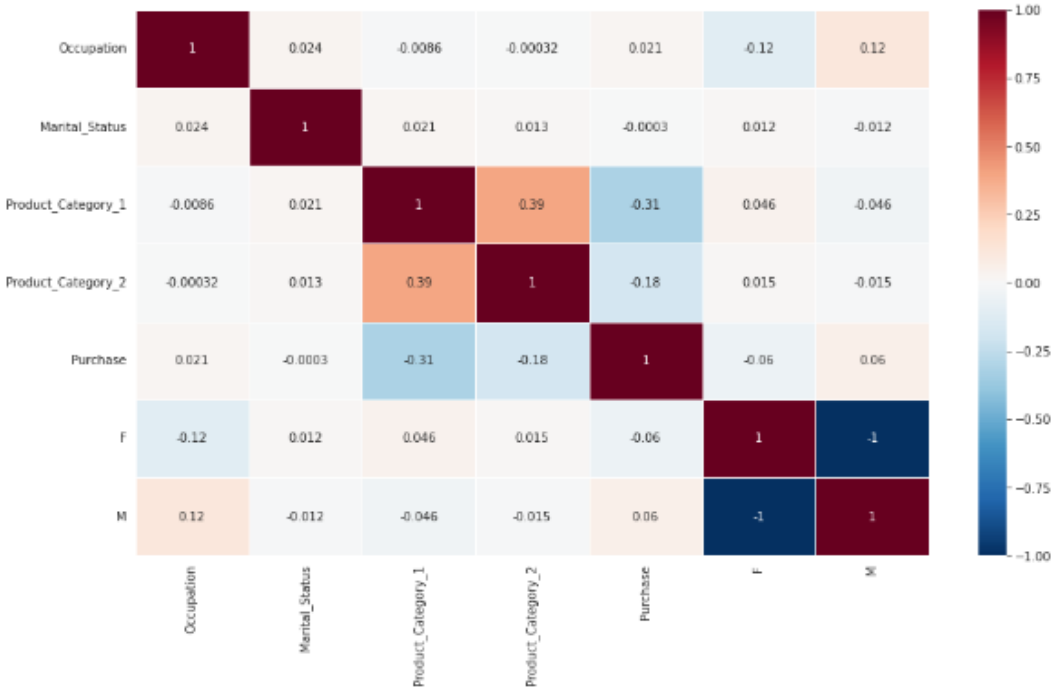
[19...

	Occupation	Marital_Status	Product_Category_1	Product_Category_2	Purchase	F	M
Occupation	1.000000	0.024287	-0.008607	-0.000318	0.021438	-0.117019	0.117019
Marital_Status	0.024287	1.000000	0.020660	0.012555	-0.000302	0.011543	-0.011543
Product_Category_1	-0.008607	0.020660	1.000000	0.393737	-0.314083	0.045827	-0.045827
Product_Category_2	-0.000318	0.012555	0.393737	1.000000	-0.181551	0.015364	-0.015364
Purchase	0.021438	-0.000302	-0.314083	-0.181551	1.000000	-0.060204	0.060204
F	-0.117019	0.011543	0.045827	0.015364	-0.060204	1.000000	-1.000000
M	0.117019	-0.011543	-0.045827	-0.015364	0.060204	-1.000000	1.000000

[192]:

```
sns.heatmap(pearsoncorr,
            xticklabels=pearsoncorr.columns,
            yticklabels=pearsoncorr.columns,
            cmap='RdBu_r',
            annot=True,
            linewidth=0.5)
```

[19... <AxesSubplot:>



Spearman Rank Correlation

[193]:
```
df['Product_Category_1'].corr(df['Product_Category_2'], method='spearman')
```

[19...
```
0.42396425892766665
```

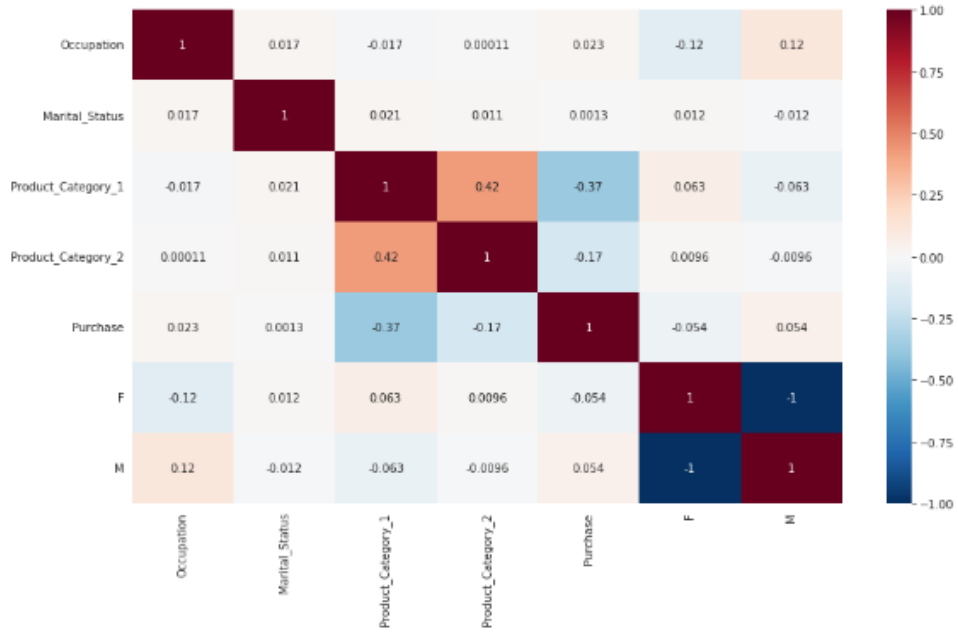
[194]:
```
corr1 = df.corr(method='spearman')
corr1
```

[19...

	Occupation	Marital_Status	Product_Category_1	Product_Category_2	Purchase	F	M
Occupation	1.000000	0.016887	-0.016549	0.000108	0.022873	-0.117082	0.117082
Marital_Status	0.016887	1.000000	0.021168	0.011121	0.001298	0.011543	-0.011543
Product_Category_1	-0.016549	0.021168	1.000000	0.423964	-0.368832	0.063480	-0.063480
Product_Category_2	0.000108	0.011121	0.423964	1.000000	-0.171937	0.009578	-0.009578
Purchase	0.022873	0.001298	-0.368832	-0.171937	1.000000	-0.054173	0.054173
F	-0.117082	0.011543	0.063480	0.009578	-0.054173	1.000000	-1.000000
M	0.117082	-0.011543	-0.063480	-0.009578	0.054173	-1.000000	1.000000

[195]:
```
plt.figure(figsize=(14,8))
sns.heatmap(corr1, annot=True, cmap='RdBu_r')
```

[19...
```
<AxesSubplot:>
```



Kendall's Rank Correlation

[198]:
```
# Data Visualisation Settings
%matplotlib inline
plt.rcParams['figure.figsize'] = 5,4
sns.set_style('whitegrid')
```

[199]:
```
corr2 = df.corr(method='kendall')
corr2
```

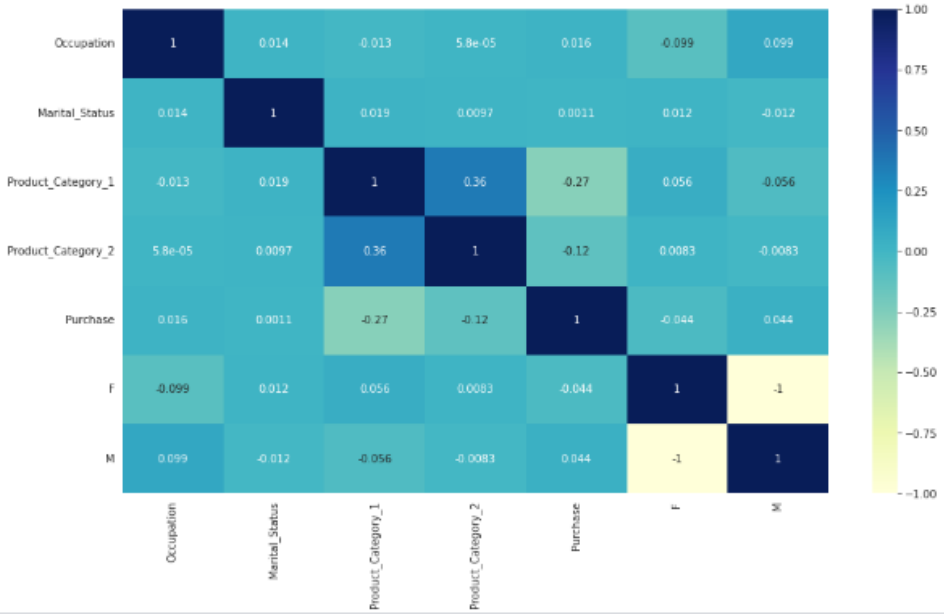
```
/opt/conda/lib/python3.7/site-packages/scipy/stats/stats.py:4812: RuntimeWarning: overflow encountered in long_scalars
  (2 * xtie * ytie) / m + x0 * y0 / (9 * m * (size - 2)))
```

[19…

	Occupation	Marital Status	Product Category 1	Product Category 2	Purchase	F	M
Occupation	1.000000	0.014295	-0.012561	0.000058	0.015803	-0.099110	0.099110
Marital Status	0.014295	1.000000	0.018789	0.009661	0.001060	0.011543	-0.011543
Product Category 1	-0.012561	0.018789	1.000000	0.360872	-0.273756	0.056345	-0.056345
Product Category 2	0.000058	0.009661	0.360872	1.000000	-0.121163	0.008320	-0.008320
Purchase	0.015803	0.001060	-0.273756	-0.121163	1.000000	-0.044235	0.044235
F	-0.099110	0.011543	0.056345	0.008320	-0.044235	1.000000	-1.000000
M	0.099110	-0.011543	-0.056345	-0.008320	0.044235	-1.000000	1.000000

[200]:
```
plt.rcParams['figure.figsize'] = 14.7,8.27
sns.heatmap(corr2,
            xticklabels=corr.columns.values,
            yticklabels=corr.columns.values,
            cmap="YlGnBu",
           annot=True)
```

[20…] <AxesSubplot:>



Chi-Squared Test

[201]:
```
contingency= pd.crosstab(df['Purchase'], df['Marital_Status'])
contingency
```

[20…]

Marital_Status	0	1
Purchase		
185	4	0
186	3	1
187	3	1
188	5	1
189	2	0
...	...	...
23956	1	0
23958	2	2
23959	1	1
23960	1	3
23961	2	1

17995 rows × 2 columns

[202]:
```
c, p, dof, expected = chi2_contingency(contingency)
print(p)
```

```
0.103804688934077
```

Conclusion-

We have successfully performed an implementation of statistical hypothesis test using scipy and sci-kit learn. We used various tests such as Pearson's Correlation Coefficient, Spearman's Rank Correlation, Kendall's Rank Correlation and Chi-Squared Test along with visualizations to find correlation in our dataset.