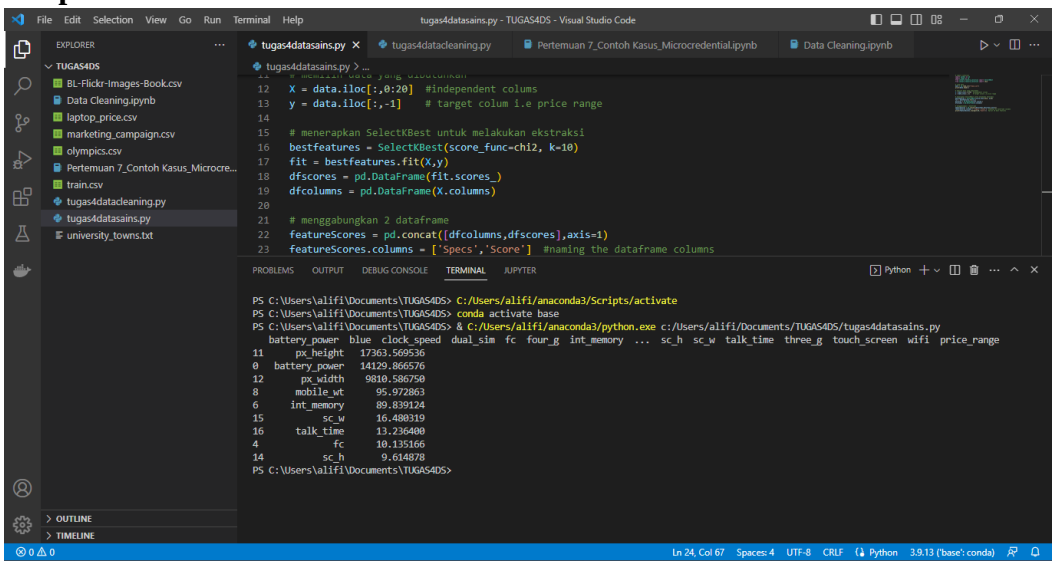
Tugas 4 Data Sains dan Analisis

1. Seleksi Unvariate

Source Code

```
# import libraries
import pandas as pd
import numpy as np
from sklearn.feature_selection import SelectKBest
from sklearn.feature_selection import chi2

# load data
data = pd.read_csv("train.csv")
print(data.head())

# memilih data yang dibutuhkan
X = data.iloc[:,0:20]  #independent colums
y = data.iloc[:,-1]    # target colum i.e price range

# menerapkan SelectKBest untuk melakukan ekstraksi
bestfeatures = SelectKBest(score_func=chi2, k=10)
fit = bestfeatures.fit(X,y)
dfscores = pd.DataFrame(fit.scores_)
dfcolumns = pd.DataFrame(X.columns)

# menggabungkan 2 dataframe
featureScores = pd.concat([dfcolumns,dfscores],axis=1)
featureScores.columns = ['Specs','Score']  #naming the dataframe columns
print(featureScores.nlargest(10,'Score'))  #print 10 best features
```

Output

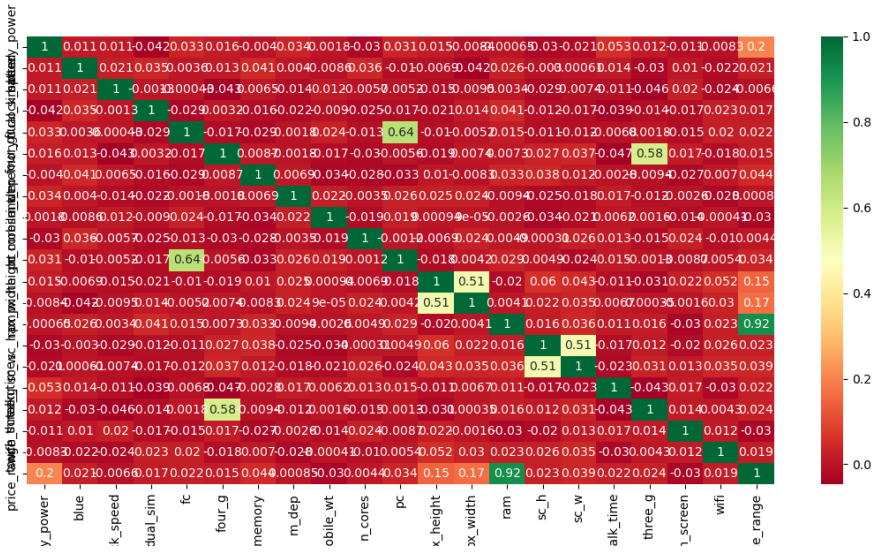


2. Matriks Kolerasi dengan Heatmap (train.csv)

Souce Code

```
# import library
import pandas as pd
import numpy as np
import seaborn as sns
import plotly.graph_objects as go
import matplotlib.pyplot as plt

# memuat data
data = pd.read_csv("train.csv")
X = data.iloc[:,0:20]  #independent columns
y = data.iloc[:,-1]    #target column i.e price range

# mendapatkan  correlations dari setiap fitur dalam dataset
corrmat = data.corr()
top_corr_features = corrmat.index

# plot heatmap
plt.figure(figsize=(20,20))
g=sns.heatmap(data[top_corr_features].corr(),annot=True,cmap="RdYlGn")
plt.show()
```

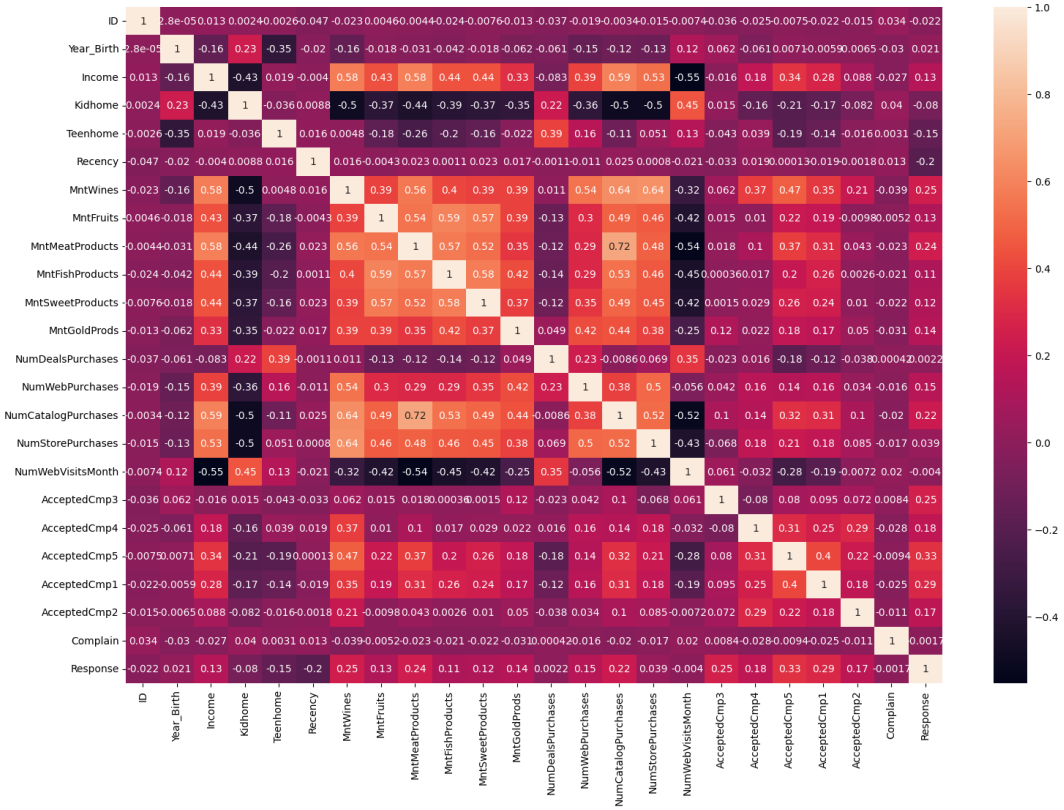
Output



Source Code (marketing_campaign.csv)

```
# import necessary libraries
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt

# load data
data = pd.read_csv("marketing_campaign.csv")

# plot heatmap of correlation matrix
plt.figure(figsize=(18,22)) # set the size of the plot
g = sns.heatmap(data.corr(), annot=True) # plot the heatmap with annotations
plt.show() # show the plot
```

Output



Data Cleanning

Source Code

```python
# import library
import pandas as pd
import numpy as np

# load data
df = pd.read_csv("laptop_price.csv")
df.head()

# periksa tipe datanya utk masing2 varibael/kolom/fitur dengan .dtypes
print(df.dtypes)

# melihat statistik dasar
df.describe()

# melihat pada lima entri pertama
print("Setelah Data di Drop\n")
print(df.head())

# memilih kolom-kolom yang akan di drop
to_drop = ['Processor','Warranty','RAM']

# melakukan drop pada data yang dipilih
df.drop(to_drop, inplace=True, axis=1)
print("Setelah Data di Drop\n")
print(df.head())
```

Output