Panda

Starter

```
import pandas as pd
!gdown 1m_uJkaKZvX24wdyAjXFIxTUY3SIawaCF
data = pd.read_csv("/content/CarPrice.csv")
data.head()
```

Describing data

```
data.describe()
data.info() # show non-null columns
data.isnull().sum()
```

Modify data

```
# drop CarName column, axis=1 means column
data.drop(['CarName'], axis = 1, inplace = True)
data.drop(columns=['B', 'C'], inplace = True)

# Insert CompanyNames in the 3rd index with CompanyName as column name
data.insert(3, "CompanyName", CompanyNames)

# make all values lowercase of CompanyName column
data.CompanyName = data.CompanyName.str.lower()

# replace values a with b in CompanyName column
data.CompanyName.replace(a, b, inplace = True)

# replace multiple columns

data.rename(columns={data.columns[17]:'very_safe',data.columns[18]:
'safe'},inplace=True)



# add suffix to every values of a column
data.doornumber = data.doornumber + "_doors"
# two becomes two_doors

# Split Data
```

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X_data, y_data,
test_size=0.20, random_state=42)

# Generate dummies

```python
def gen_dummies(column_name, df):
    temp = pd.get_dummies(df[column_name])
    df = pd.concat([df, temp], axis = 1)
    df.drop([column_name], axis = 1, inplace = True)
    return df

# generate dummies for fueltype column
# delete fueltype column
# create new columns like gas, diesel
# put 0/1 in those columns
data = gen_dummies('fueltype', data)
```

Cleaning data

Remove empty cells

```
data = data.dropna()
data.dropna(inPlace = True)

# removing from only date column
df.dropna(subset=['Date'], inplace = True)
```

Replace empty cells

```
data.fillna(-1, inplace = True)
data['col'].fillna(-1, inplace = True) # only clean for col

# replacing with mean
x = df["Calories"].mean() # or .median()
df["Calories"].fillna(x, inplace = True)

# replacing the value of 'Duration' column in the 7th row
df.loc[7, 'Duration'] = 45
```

Loop through all values in the "Duration" column.
If the value is higher than 120, set it to 120:
```

```
for x in df.index:
  if df.loc[x, "Duration"] > 120:
    df.loc[x, "Duration"] = 120
```

Remove duplicates

```
df.drop_duplicates(inplace = True)
```

Corelation

```
df.corr()
```