Team Number: 7

Team Captain: Jacob Silva

Team Members: Juliana Steele

Joel Hurtado

Problem 1.1

```
import pandas as pd

df = pd.read_excel('House_Prices_PRED.xlsx')
```

Problem 1.2

```
import numpy as np
from sklearn.model_selection import train_test_split

actual_price = df['SalePrice']
predicted_price=df['SalePrice_MP']

residual = (actual_price - predicted_price)
squared_errors = (residual) ** 2


SSE = squared_errors.sum()


# Calculate the number of data points
n = len(actual_price)

# Calculate Average Squared Error (ASE)
ASE = SSE / n


print(f"Sume Squared error is: ", {SSE})
print(f"")
print(f"Average Squared error is: ", {ASE})
```

```
Sume Squared error is:  {968603985509.3241}

Average Squared error is:  {663427387.3351535}
```

Problem 1.3

```
mean_y = sum(actual_price)/n

SS = (actual_price - mean_y)**2

TSS = SS.sum()
r_squared = 1-(SSE/TSS)
print(f"The r^2 value is: ",{r_squared})
```

```
The r^2 value is:  {0.8948074161109033}
```

Problem 1.4

```
ABS_div = abs(residual / actual_price)
M1 = ABS_div.sum()
MAPE = (1/n)*M1
print(f"MAPE: ",{MAPE})
```

```
MAPE:  {0.0814515526875231}
```

Problem 1.5

```
ABS = abs(residual)
M2 = ABS.sum()
MAE = (1/n)*M2
print(f"MAE: ", {MAE})
```

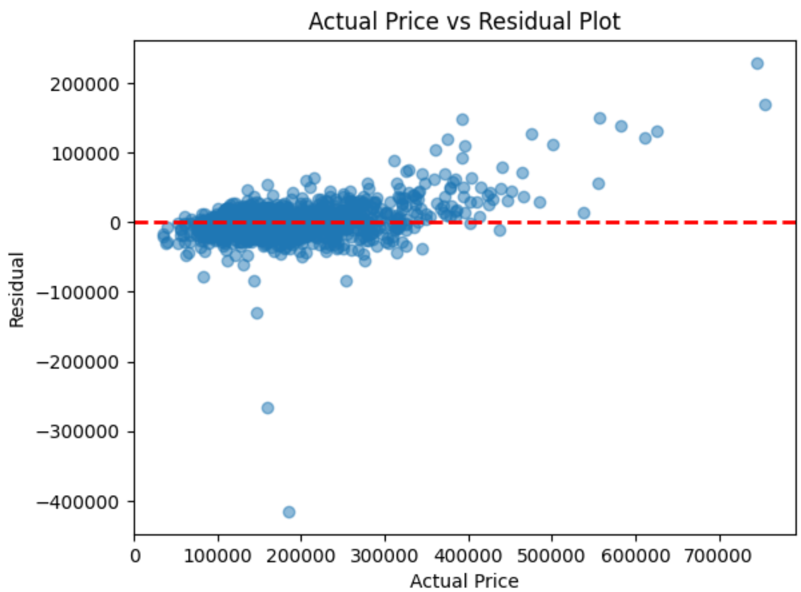
```
MAE:  {14368.025828767124}
```

Problem 1.6

```
import matplotlib.pyplot as plt
df['Residuals'] = df['SalePrice'] - df['SalePrice_MP']
print(df.head())
X = df['SalePrice']
y= df['Residuals']

plt.scatter(X, y, alpha=0.5)
plt.axhline(y=0, color='r', linestyle='--', linewidth=2)
plt.xlabel('Actual Price')
plt.ylabel('Residual')
plt.title('Actual Price vs Residual Plot')
plt.show()
```

	Id	SalePrice	SalePrice_MP	Residuals
0	1	208500	207439.62	1060.38
1	2	181500	174829.19	6670.81
2	3	223500	219431.19	4068.81
3	4	140000	167653.84	-27653.84
4	5	250000	282350.02	-32350.02



*Problem 2 Residual Plot for Predictors *

Problem 2.1

```
file_path = 'House_Prices_PRED.xlsx'

sheet_name = 'PB3'

df2 = pd.read_excel(file_path, sheet_name=sheet_name)

print(df2)
```

	Id	SalePrice	SalePrice_MP	MSZoning	FireplaceQu	GAR	BATH	Age	TSF
0	1	208500	207439.62	RL	NaN	548	2.5	5	1710
1	2	181500	174829.19	RL	TA	460	2.0	31	1262
2	3	223500	219431.19	RL	TA	608	2.5	7	1786
3	4	140000	167653.84	RL	Gd	642	1.0	91	1717
4	5	250000	282350.02	RL	TA	836	2.5	8	2198

```
...   ...        ...         ...       ...          ...  ...   ...  ...   ...
1455  1456     175000    171836.32       RL           TA  460   2.5    8  1647
1456  1457     210000    207697.04       RL           TA  500   2.0   32  2073
1457  1458     266500    253549.21       RL           Gd  252   2.0   69  2340
1458  1459     142125    143850.93       RL          NaN  240   1.0   60  1078
1459  1460     147500    152684.65       RL          NaN  276   1.5   43  1256

[1460 rows x 9 columns]
```

Problem 2.2

```
 import numpy as np
from sklearn.model_selection import train_test_split

actual_price = df2['SalePrice']
predicted_price=df2['SalePrice_MP']

residual = (actual_price - predicted_price)
```

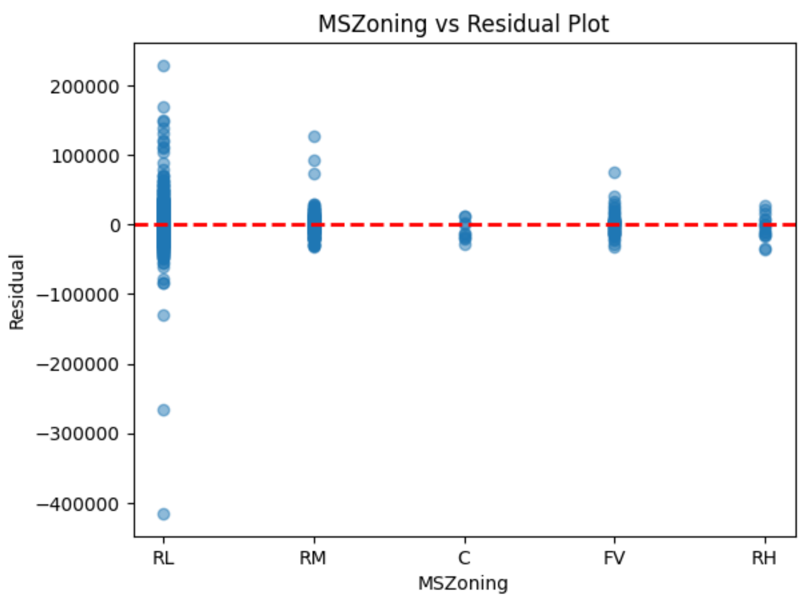
Problem 2.3

```
import matplotlib.pyplot as plt
df2['Residuals'] = df2['SalePrice'] - df2['SalePrice_MP']
print(df2.head())
X = df2['MSZoning']
y= df2['Residuals']

plt.scatter(X, y, alpha=0.5)
plt.axhline(y=0, color='r', linestyle='--', linewidth=2)
plt.xlabel('MSZoning')
plt.ylabel('Residual')
plt.title('MSZoning vs Residual Plot')
plt.show()
```

```
   Id  SalePrice  SalePrice_MP MSZoning FireplaceQu  GAR  BATH  Age   TSF  \
0   1     208500     207439.62       RL         NaN  548   2.5    5  1710
1   2     181500     174829.19       RL          TA  460   2.0   31  1262
2   3     223500     219431.19       RL          TA  608   2.5    7  1786
3   4     140000     167653.84       RL          Gd  642   1.0   91  1717
4   5     250000     282350.02       RL          TA  836   2.5    8  2198

   Residuals
0    1060.38
1    6670.81
2    4068.81
3  -27653.84
4  -32350.02
```

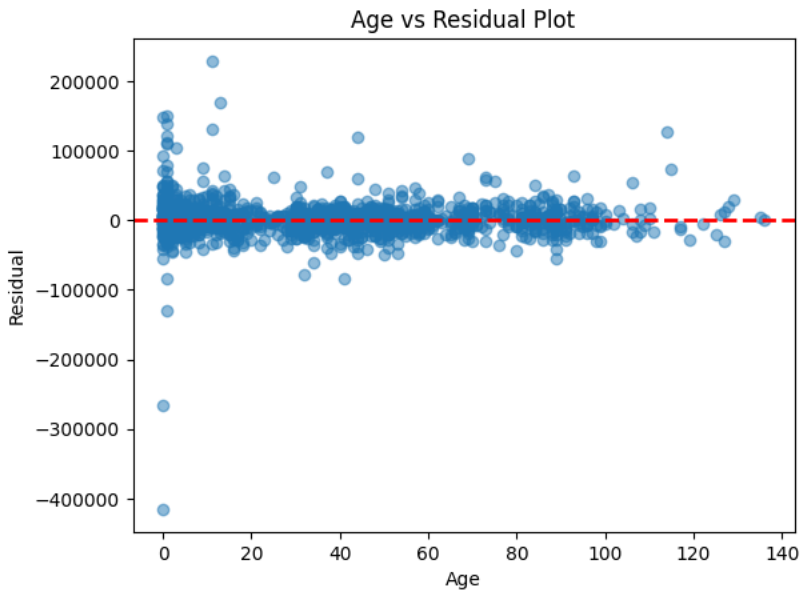


Based on the residual plot above, RL has a few outliers but the other zones the residuals seem simliar and relatively close to 0 which would be good for linear regression

Problem 2.4

```
X = df2['Age']
y= df2['Residuals']

plt.scatter(X, y, alpha=0.5)
plt.axhline(y=0, color='r', linestyle='--', linewidth=2)
plt.xlabel('Age')
plt.ylabel('Residual')
plt.title('Age vs Residual Plot')
plt.show()
```



Since the age is densely packed where the residual = 0 we assume there is heteroscedasticity, so based on the plot we can assume non constant variance

PART II: True or False

1. Suppose that multiple models were built using the same data and the MSE of these models were calculated using the training data sample, the model with the lowest MSE is the best model.

TRUE

2. We can calculate both bias and variance for any fitted model and then combine them together to get MSE.

FALSE

3. The R2 of the best model for data set A is 0.92. You built a model use data set B and the R2 of your model is 0.94. This means that your model is the best model for data set B.

FALSE

4. MSE (Mean Squared Error) calculated using the training data sample is a monotone decreasing function of the "model complexity".

TRUE

5. R^2 (R-Square) calculated using the training data is a monotone decreasing function of the "model complexity".

FALSE

6. MAPE (Mean Absolute Percentage Error) calculated using the training data is a monotone decreasing function of the “model complexity”.

FALSE

7. K-Ford cross validation method is an honest modeling error assessment methodology that should be used when the training sample size is extremely small. For example, if the training sample size is 20, we can use 5 ford cross validation.

FALSE

8. LOOCV (N-Ford Cross Validation) can be used to estimate the model error even if the training sample size is very large since it is the most efficient cross validation method.

FALSE

PART III ESSAY

a) What is the chance that jth observation is not the first observation selected into the bootstrapping sample?

The probability of the jth observation is the not first observation selected is 1 - 1/n where n is the number of observations, so its 0.995

b) What is the chance that jth observation is not the second observation selected into the bootstrapping sample?

The probability of the jth observation is the not second observation selected is the same as a), 0.995, because of the sampling with replacement.

c) What is the chance that jth observation is not the last observation selected into the bootstrapping sample?

0.995

d) What is the chance that jth observation is not in the bootstrapping sample of size 200?

Due to sampling with replacement the chances are quite low. The formula is (1 - 1/n)^n = (1- 1/200)^200, which is 0.3670

e) What is the chance that jth observation is not selected into the bootstrapping sample of size infinitely from a set of infinites many observations?

lim n→∞ (1 - 1/n)^n = 1/e = 0.37