Car Purchase

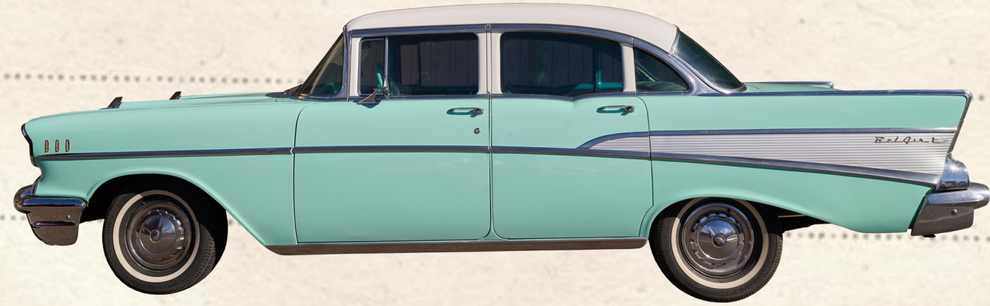
Team Alpha

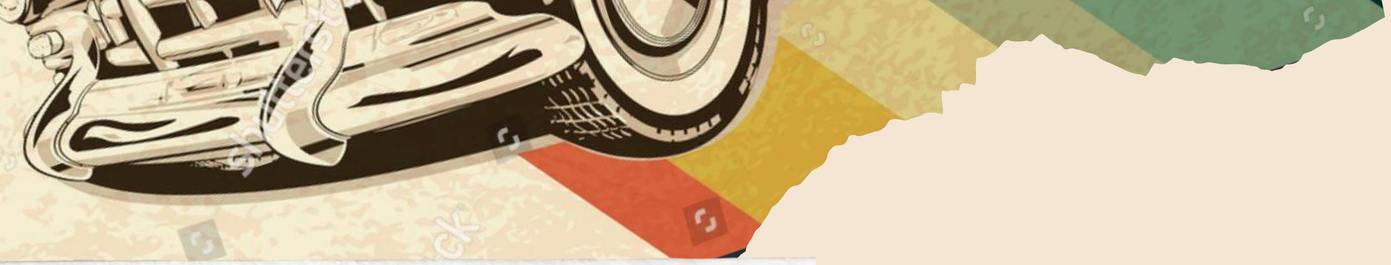
Salguerro

Lake Resort

Car Purchase

Alpha





Problem statement

Developing model

We are working as a car salesman and we would like to develop a model to predict the total dollar amount that customers are willing to pay given the following attributes:

- Customer Name
- Customer e-mail
- Country
- Gender
- Age
- Annual Salary
- Credit Card Debt
- Net Worth

The model should predict:

- Car Purchase Amount

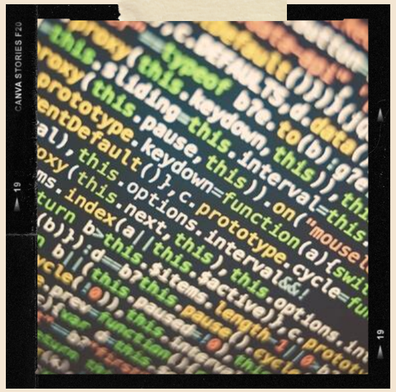
Data Exploration

Dataset

9 columns
500 lines

Numerical Variables

- Age
- Annual Salary
- Credit Card Debt
- Net Worth
- Car Purchase Amount



Categorical Variables

- Customer Name
- Customer e-mail
- Country
- Gender

Data Exploration

500 people

20 - 70 years
average age 46 years

239 women

243 men

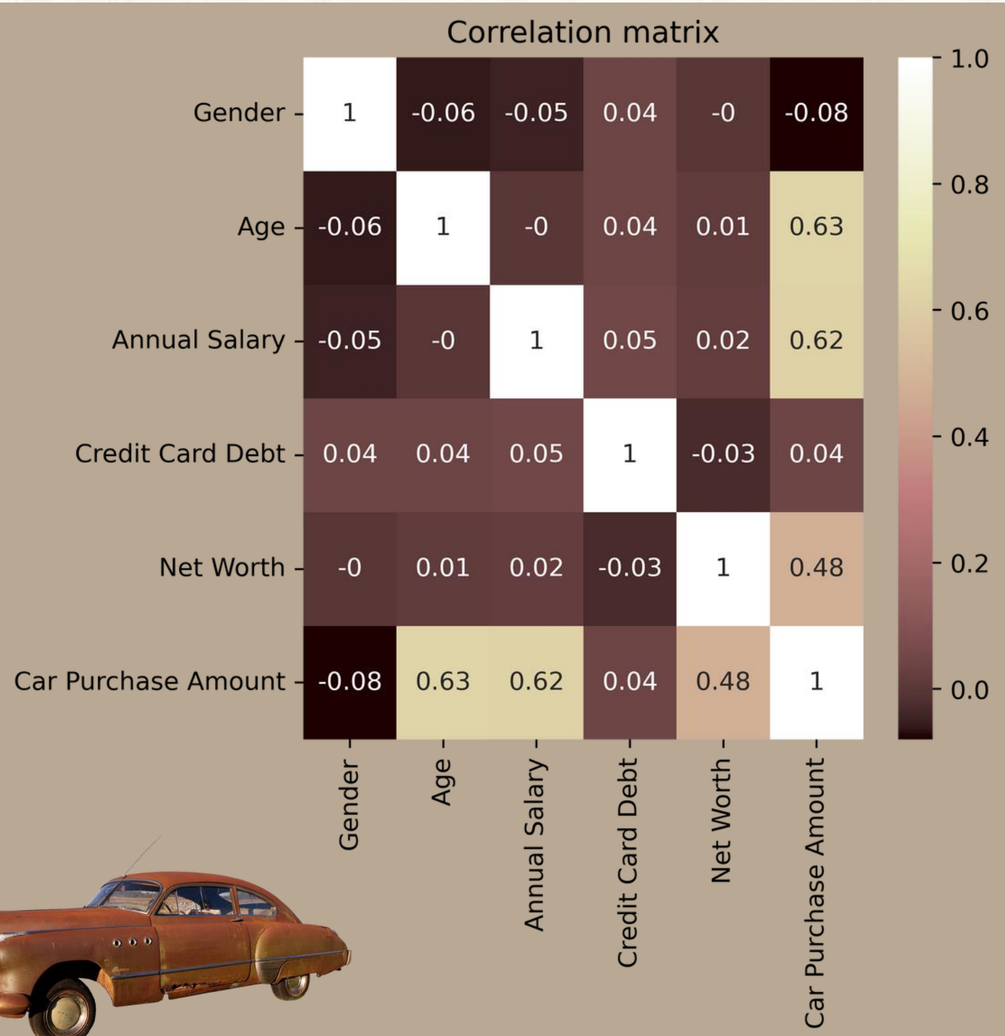
Annual Salary from $20,000 to $100,000
Average Annual Salary ~ $62,000

Net Worth from $20,000 to $100,000
Average Net Worth ~ $43,000

Credit Card Debt from $100 to $20,000
Average Credit Card Debt ~ $9,700

Car Purchase Amount from $9,000 to $80,000
Average Car Purchase Amount ~ $44,200

Numerical variables



Car Purchase Amount

is affected by variables such as

Age

0.63

Annual Salary

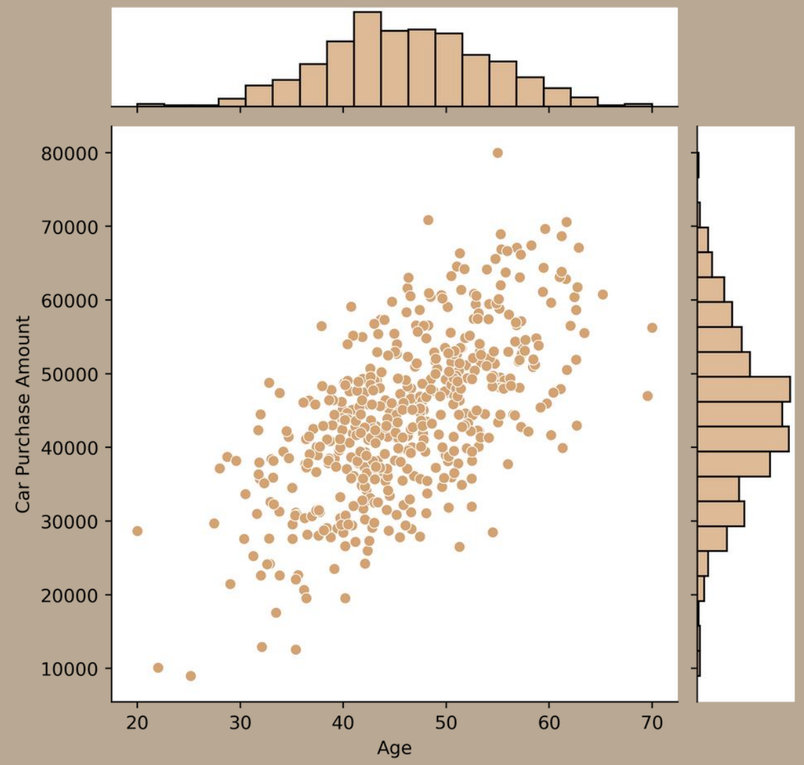
0.62

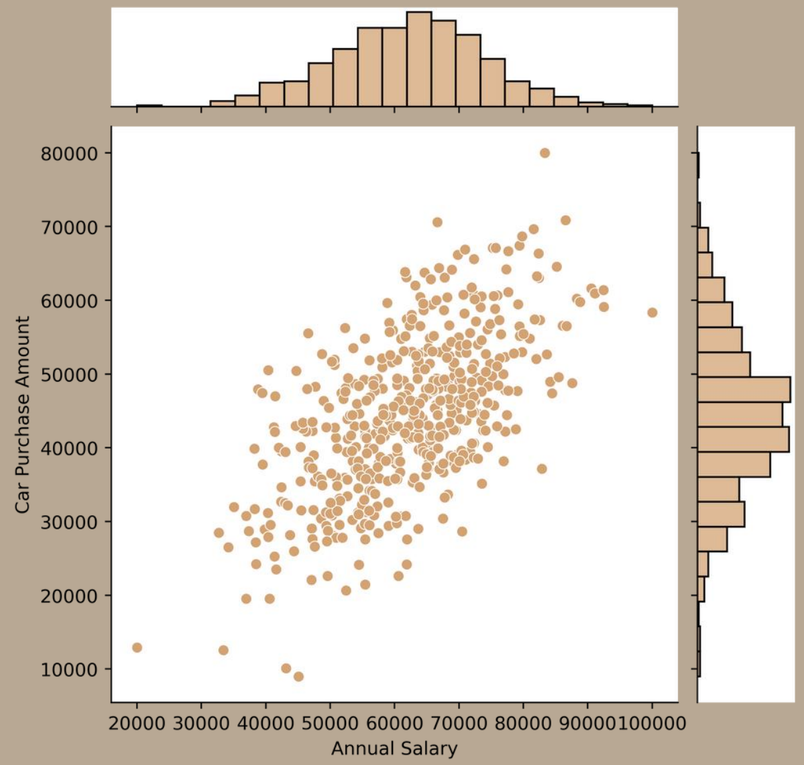
Net Worth

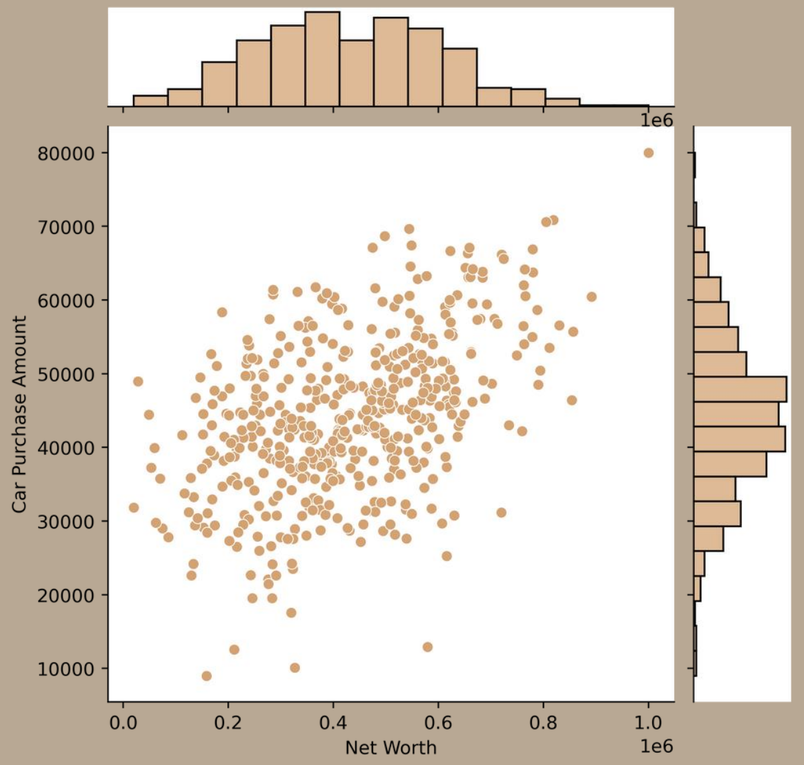
0.48

Others

have very little correlation







Age

People aged 35 to 60 are more likely to buy cars. The older the person, the more expensive car he buys.

Min $20,000

Max $100,000

Mean ~ $62,000

Annual Salary

If the annual income exceeds $45,000, then the probability that he will buy a car is high. The higher the income of a person, the more expensive the car he buys.

Min $20,000

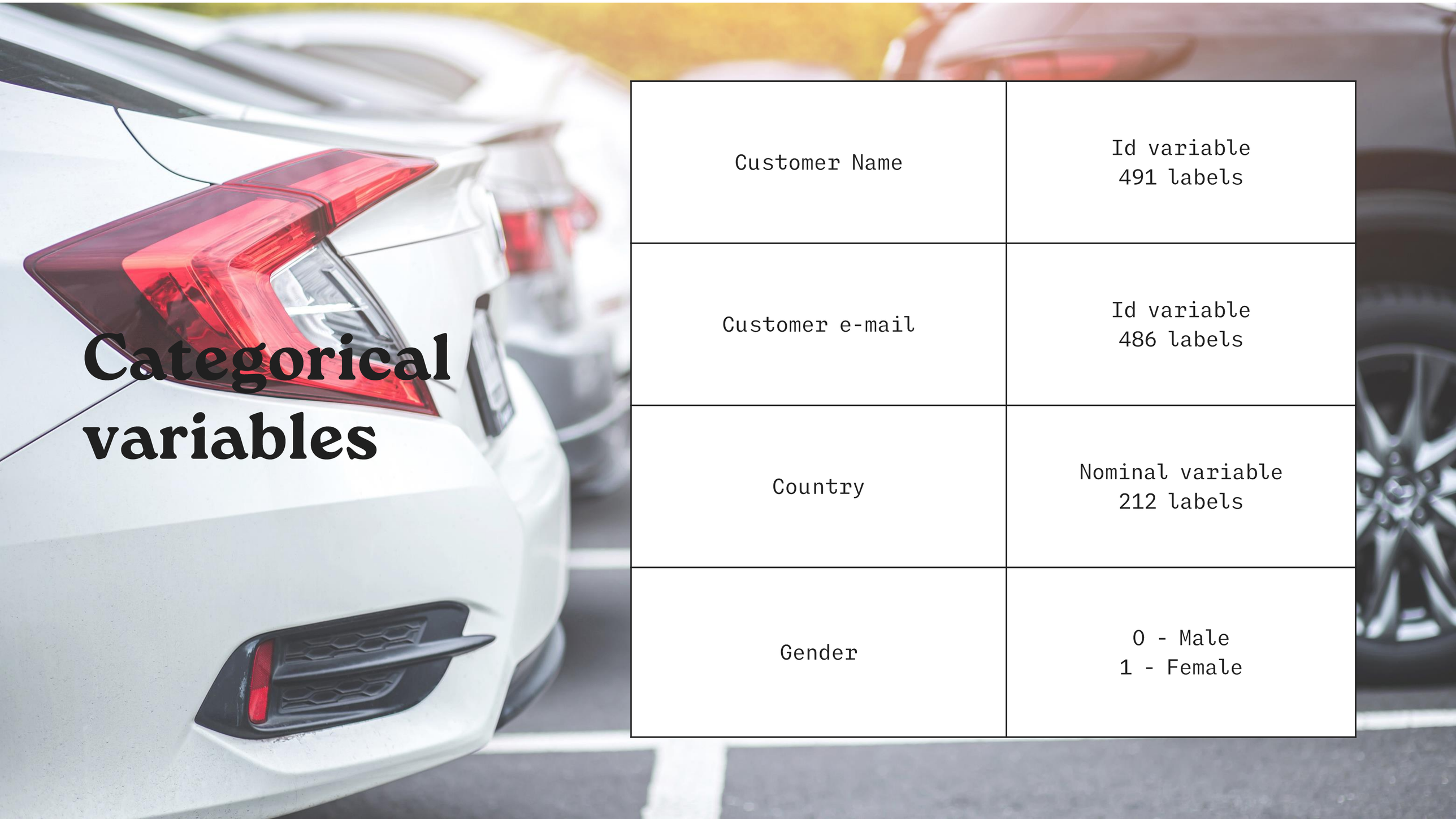
Max $100,000

Mean ~ $62,000

Net Worth

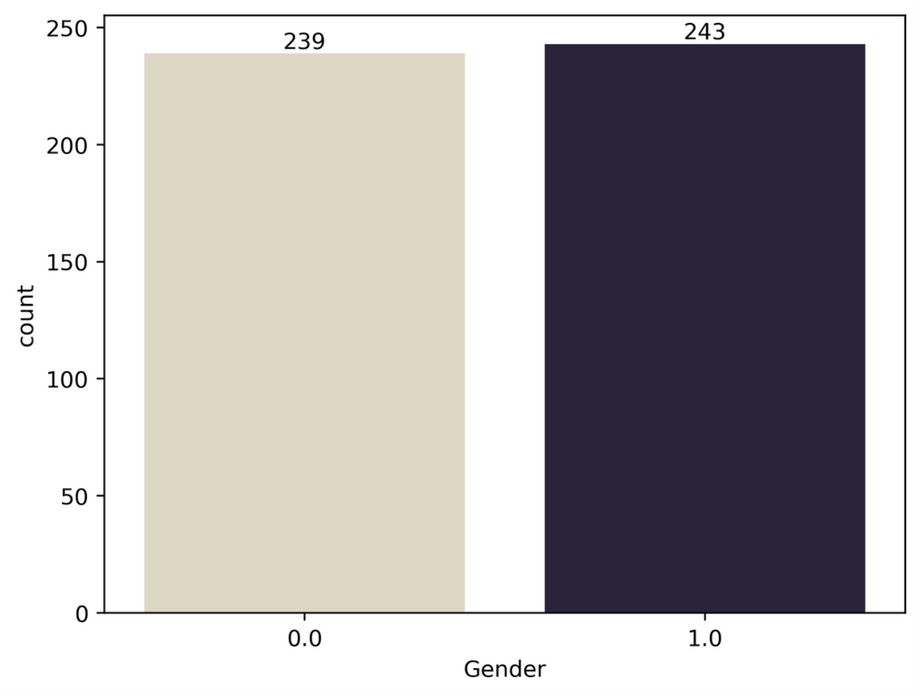
Net worth = 1 means that a man needs to have at least $10,374,030 to be in the top 1% of the U.S. Forbes, July, 2022.

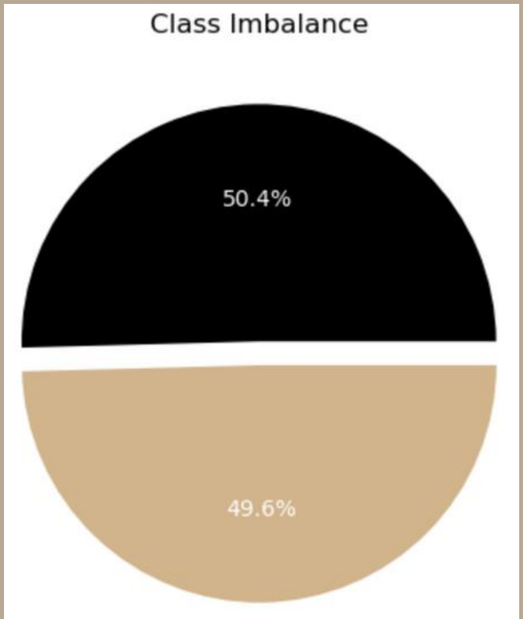
People with net worth above 0.2 are more likely to buy cars. The higher the wealth, the more expensive the car they can afford. But people with wealth above 0.6 are less likely to buy a car for themselves, and if they buy, then an expensive one.



Categorical variables

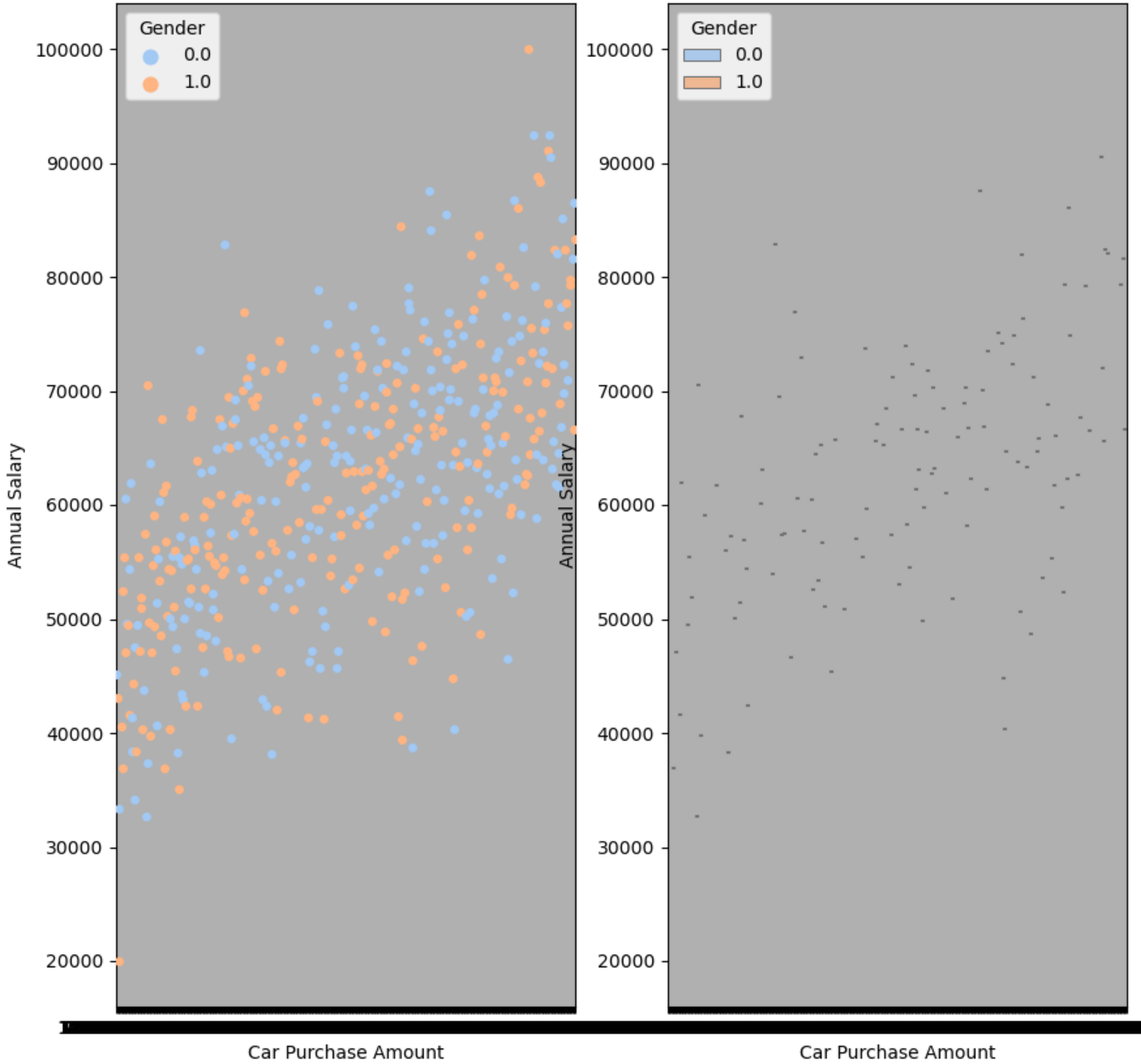
Customer Name	Id variable 491 labels
Customer e-mail	Id variable 486 labels
Country	Nominal variable 212 labels
Gender	0 - Male 1 - Female



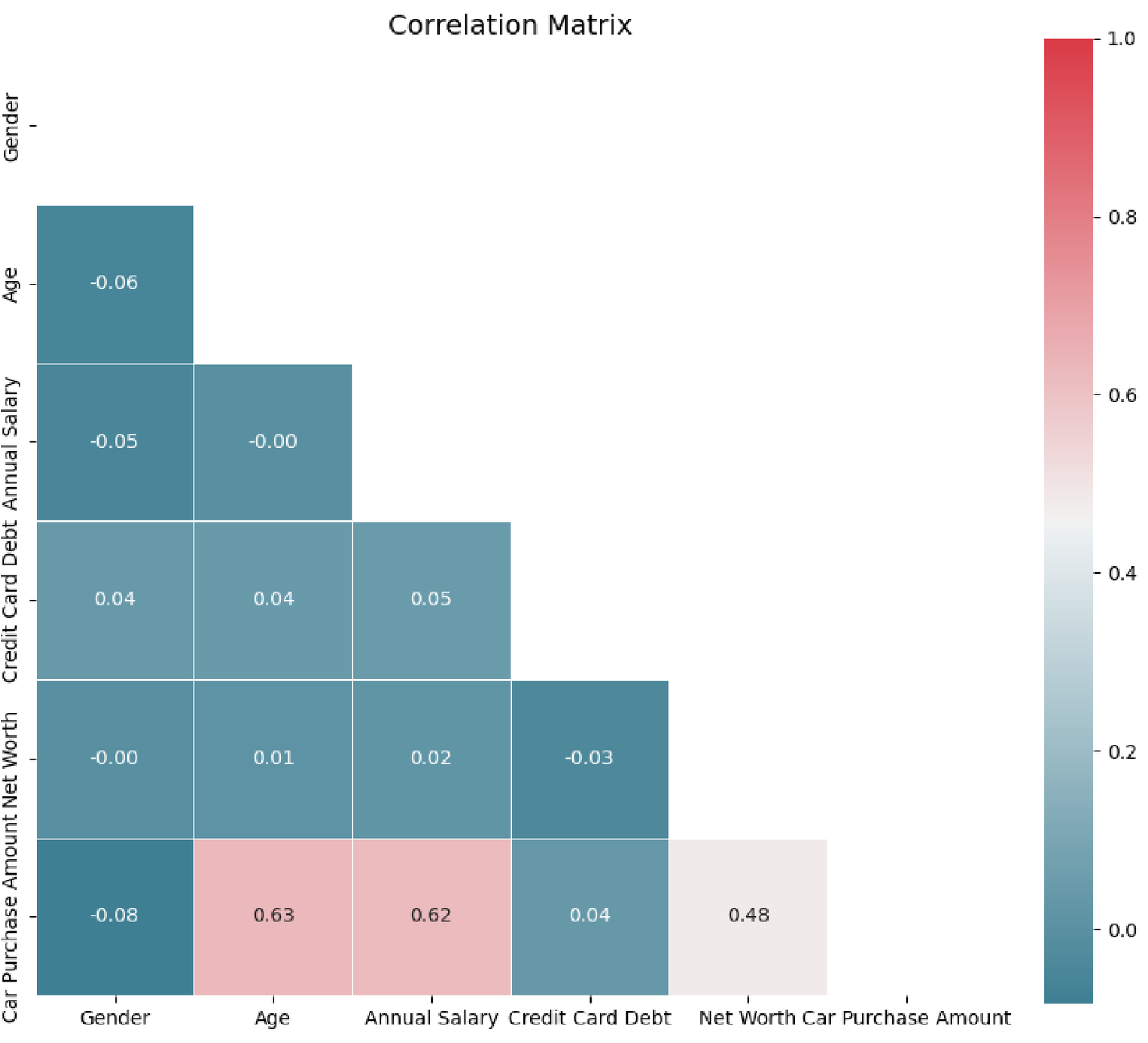


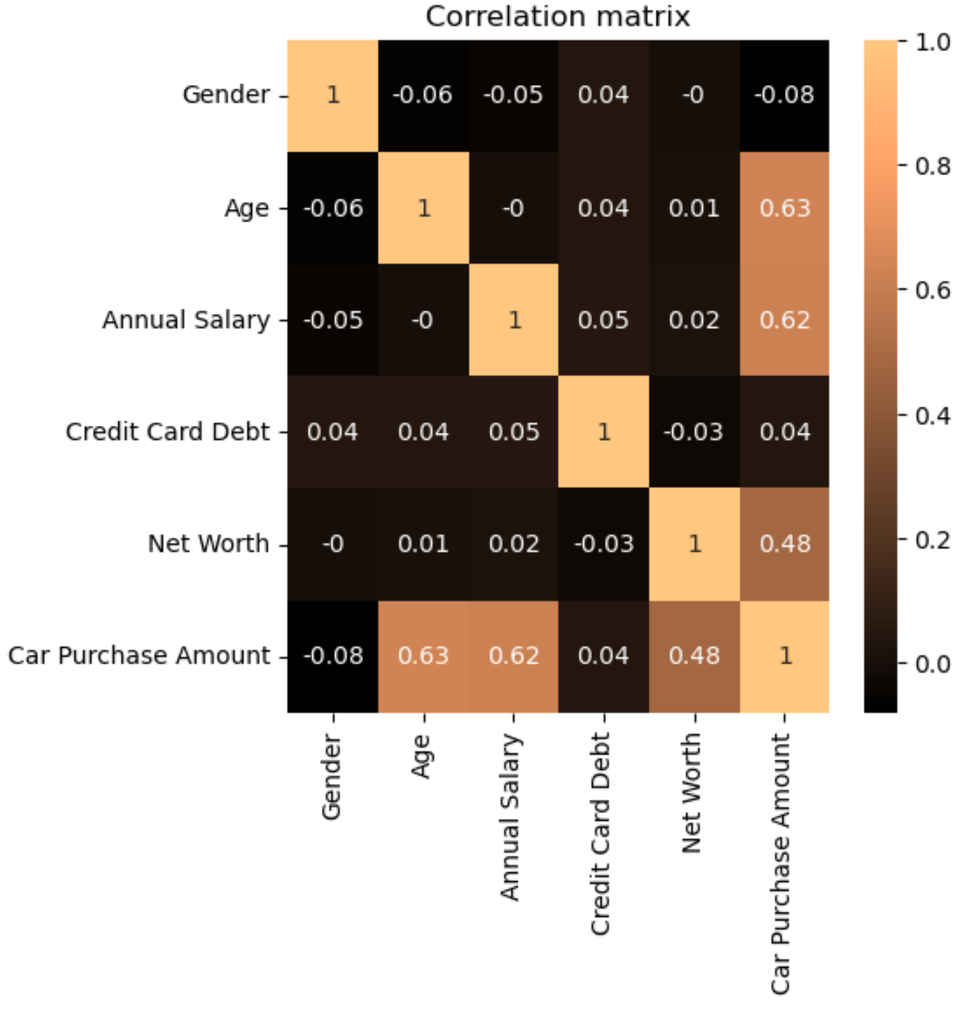
Gender

There is almost the same number of women and men participate in the dataset. Annual salary does not depend on gender. Also, car purchase amount too does not depend on gender.



Correlation matrix different styles

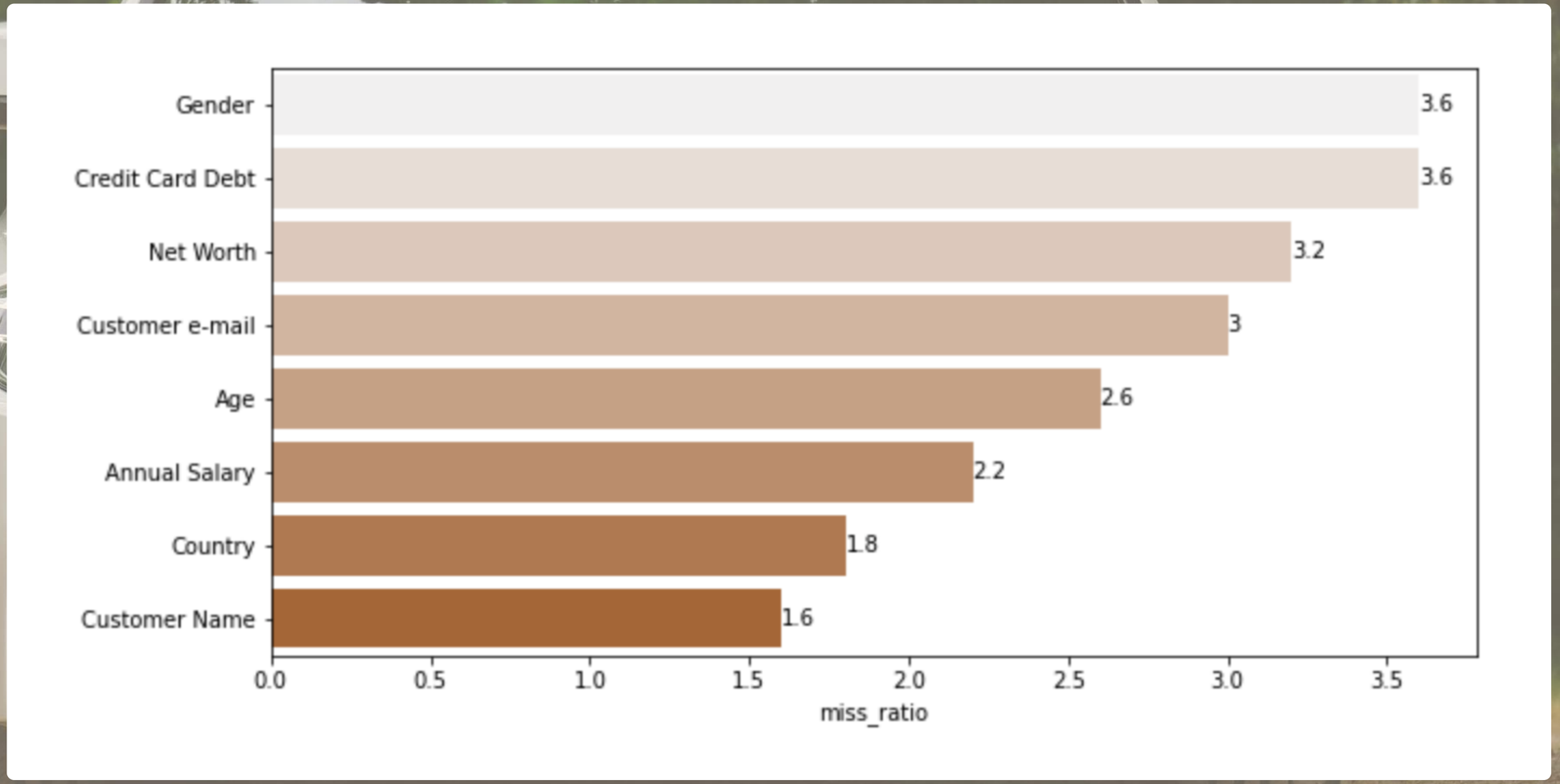


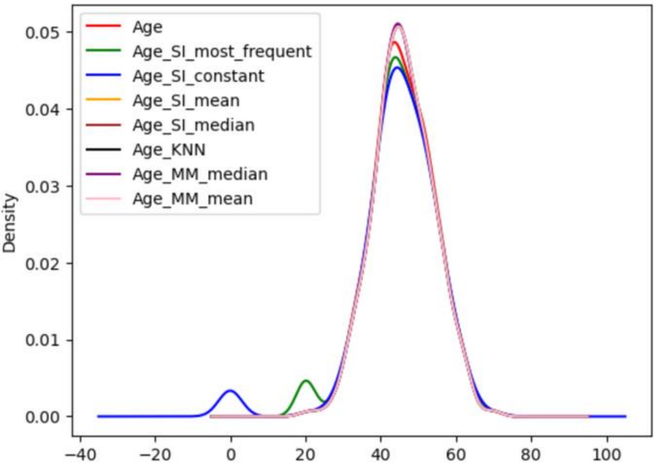


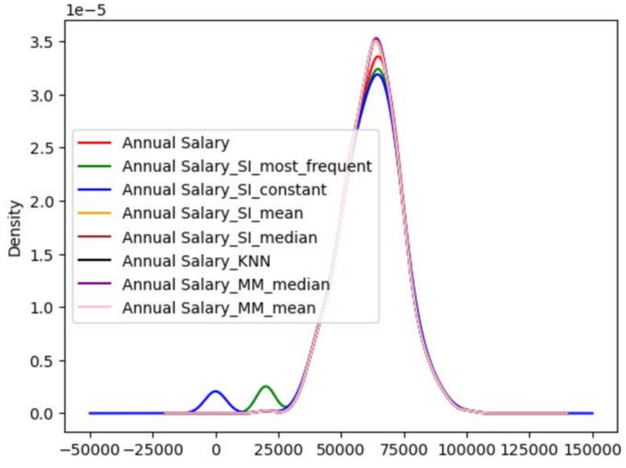
Imputation of missing values

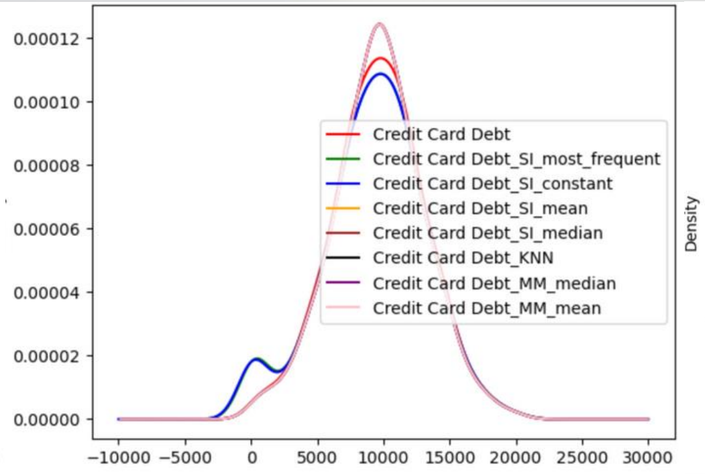
Car Purchase Amount has no missing values.

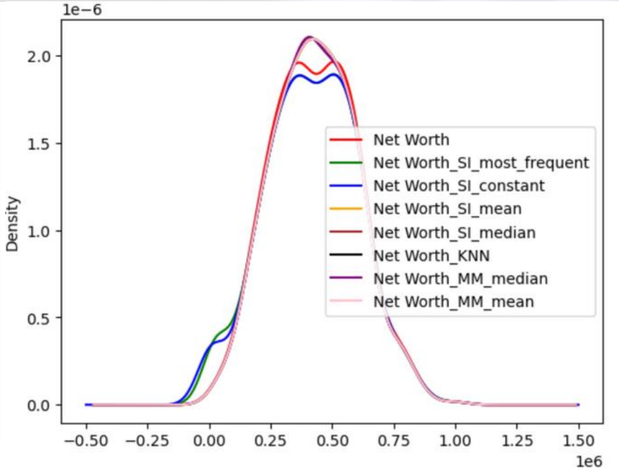
All variables have a small percentage of missing data.

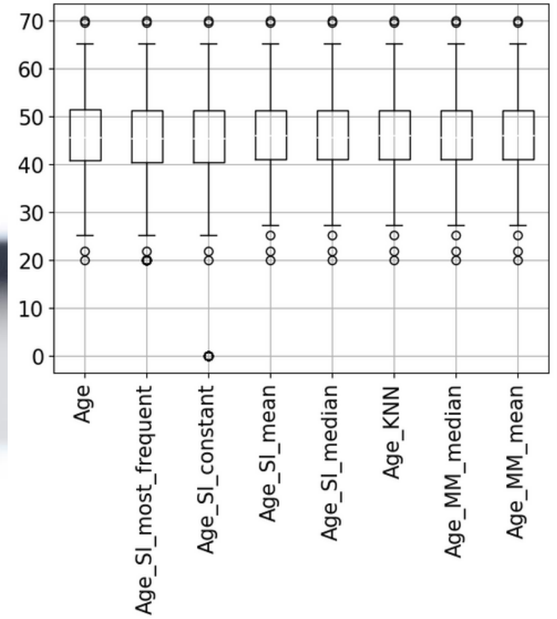


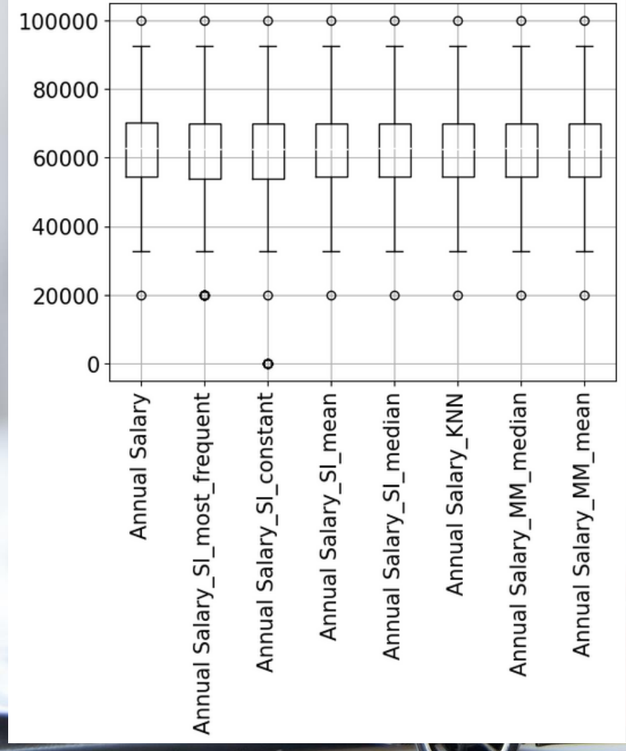

Age
Age_SI_most_frequent
Age_SI_constant
Age_SI_mean
Age_SI_median
Age_KNN
Age_MM_median
Age_MM_mean
Density

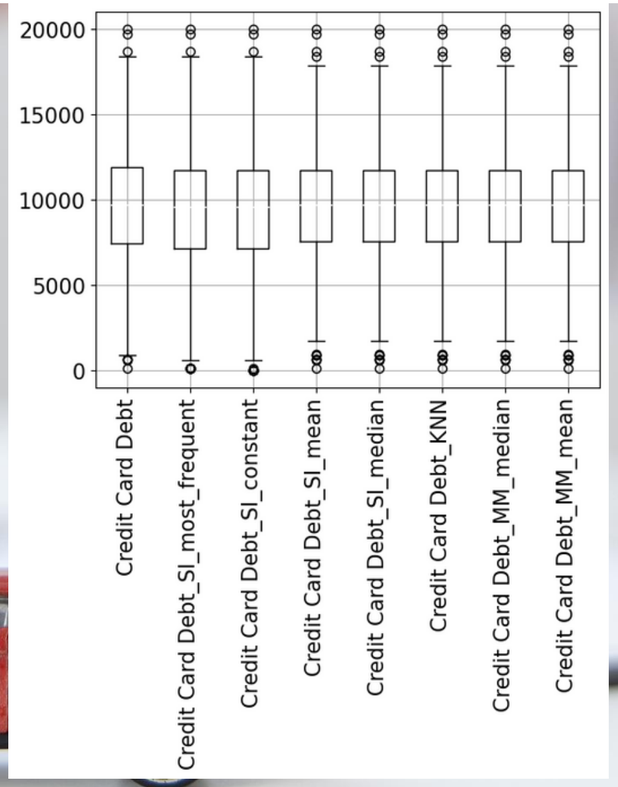

1e−5
Annual Salary
Annual Salary_SI_most_frequent
Annual Salary_SI_constant
Annual Salary_SI_mean
Annual Salary_SI_median
Annual Salary_KNN
Annual Salary_MM_median
Annual Salary_MM_mean
Density

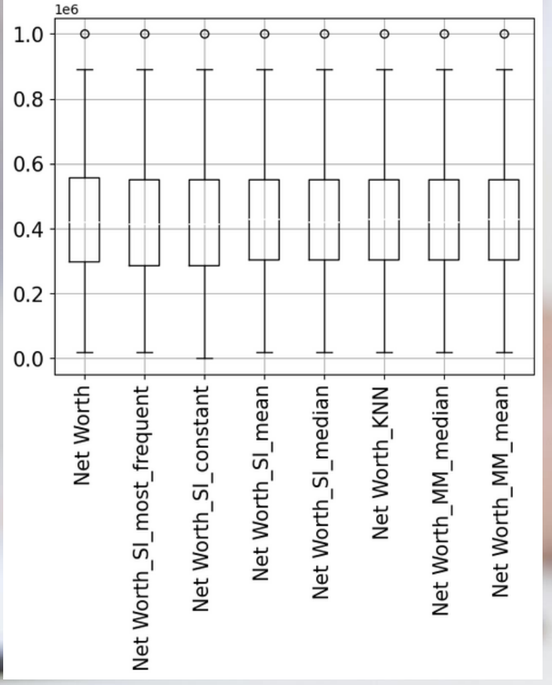

Credit Card Debt
Credit Card Debt_SI_most_frequent
Credit Card Debt_SI_constant
Credit Card Debt_SI_mean
Credit Card Debt_SI_median
Credit Card Debt_KNN
Credit Card Debt_MM_median
Credit Card Debt_MM_mean

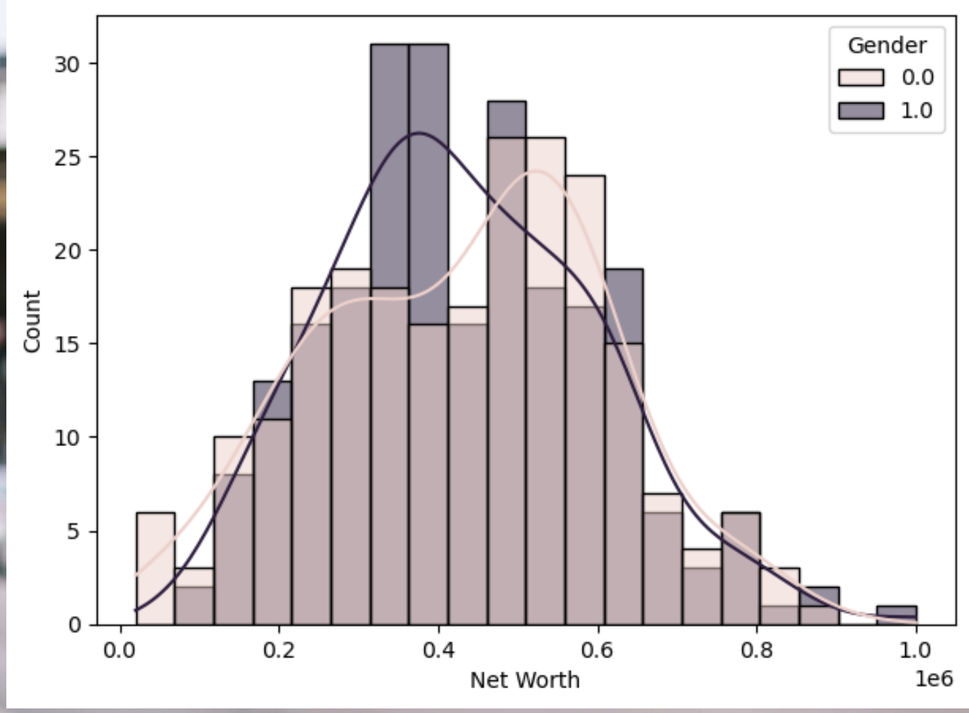

1e−6
Net Worth
Net Worth_SI_most_frequent
Net Worth_SI_constant
Net Worth_SI_mean
Net Worth_SI_median
Net Worth_KNN
Net Worth_MM_median
Net Worth_MM_mean
Density
1e6

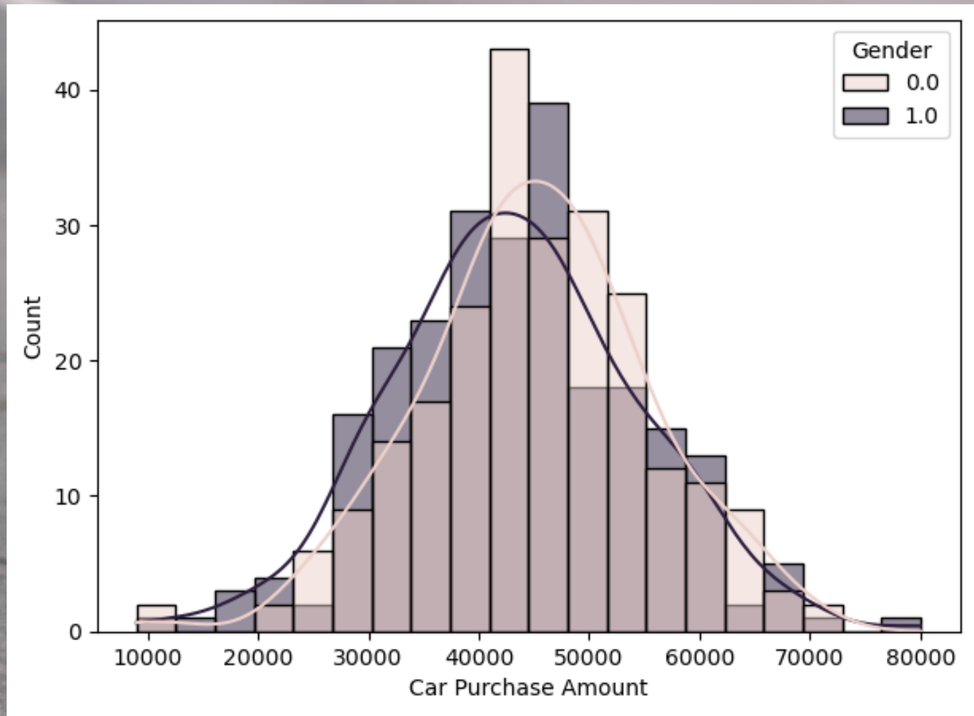

Age
Age_SI_most_frequent
Age_SI_constant
Age_SI_mean
Age_SI_median
Age_KNN
Age_MM_median
Age_MM_mean

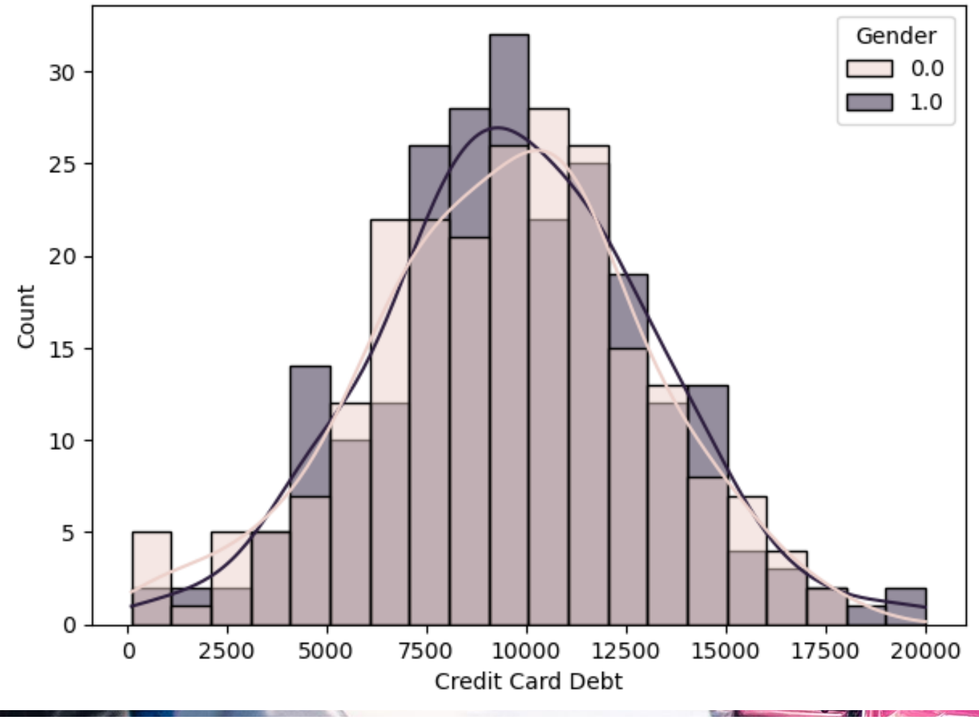

Annual Salary
Annual Salary_SI_most_frequent
Annual Salary_SI_constant
Annual Salary_SI_mean
Annual Salary_SI_median
Annual Salary_KNN
Annual Salary_MM_median
Annual Salary_MM_mean


Credit Card Debt
Credit Card Debt_SI_most_frequent
Credit Card Debt_SI_constant
Credit Card Debt_SI_mean
Credit Card Debt_SI_median
Credit Card Debt_KNN
Credit Card Debt_MM_median
Credit Card Debt_MM_mean

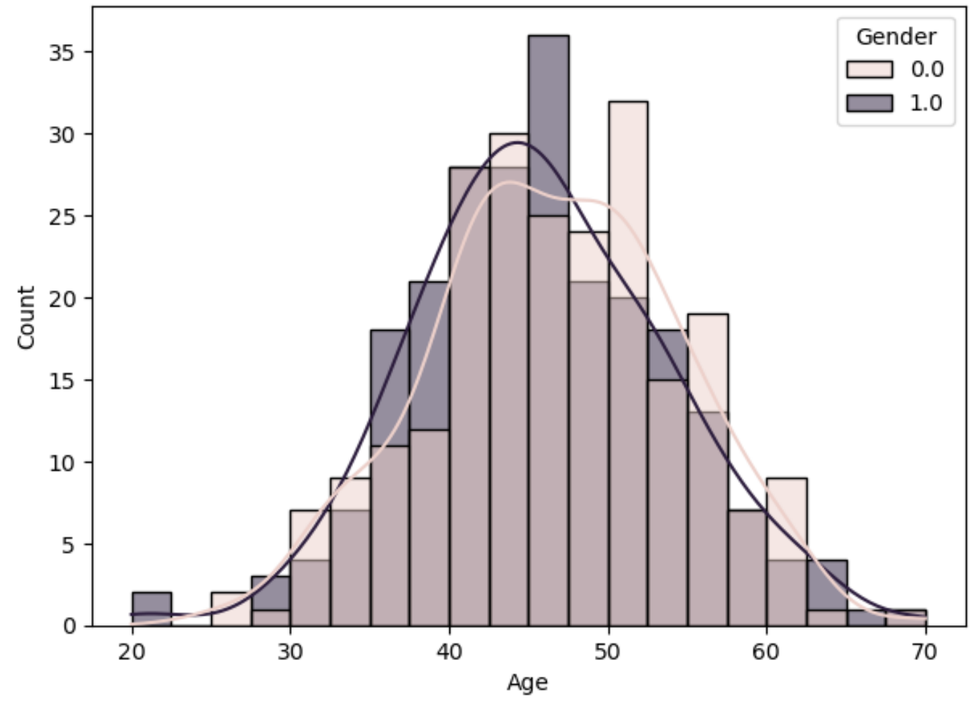

1e6
Net Worth
Net Worth_SI_most_frequent
Net Worth_SI_constant
Net Worth_SI_mean
Net Worth_SI_median
Net Worth_KNN
Net Worth_MM_median
Net Worth_MM_mean

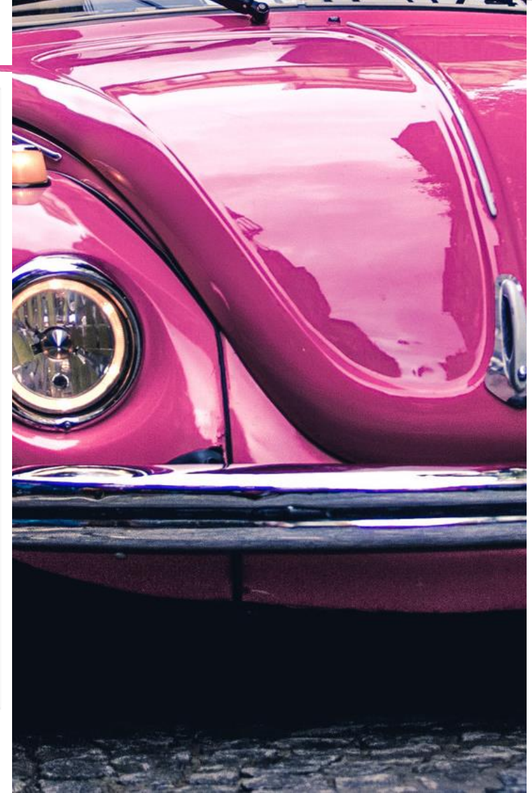

Gender
0.0
1.0
Count
Net Worth
1e6

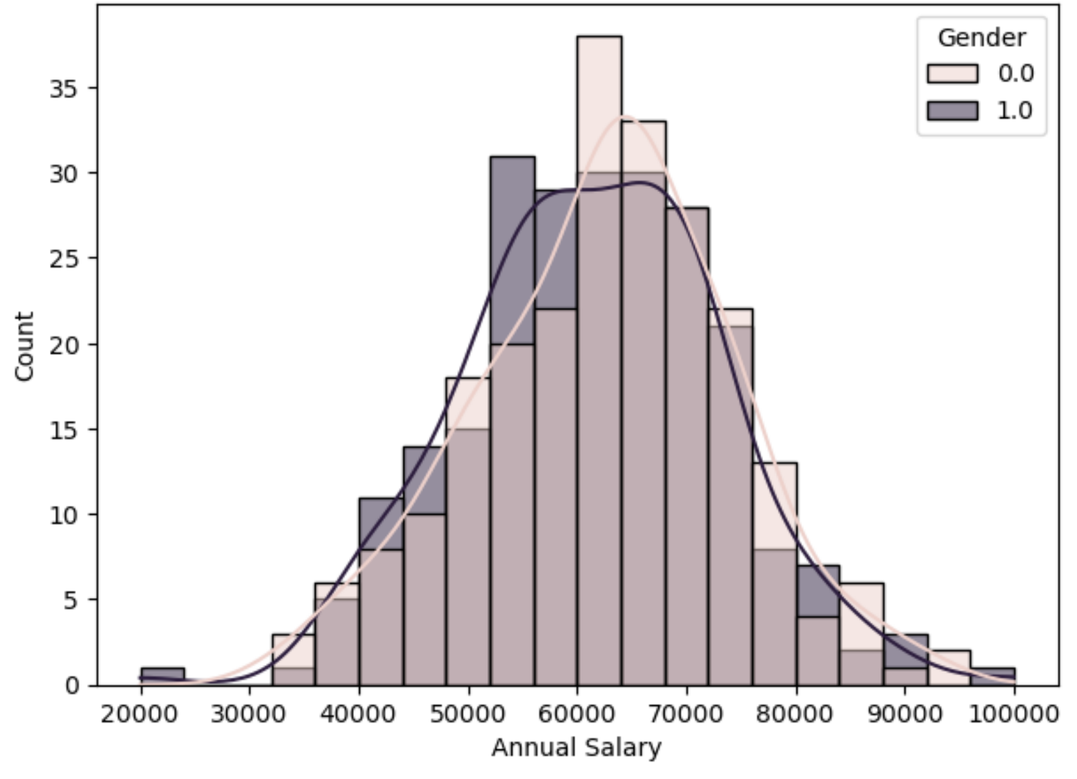

Gender
0.0
1.0
Count
Car Purchase Amount


Gender
0.0
1.0
Count
Credit Card Debt
0
2500
5000
7500
10000
12500
15000
17500
20000

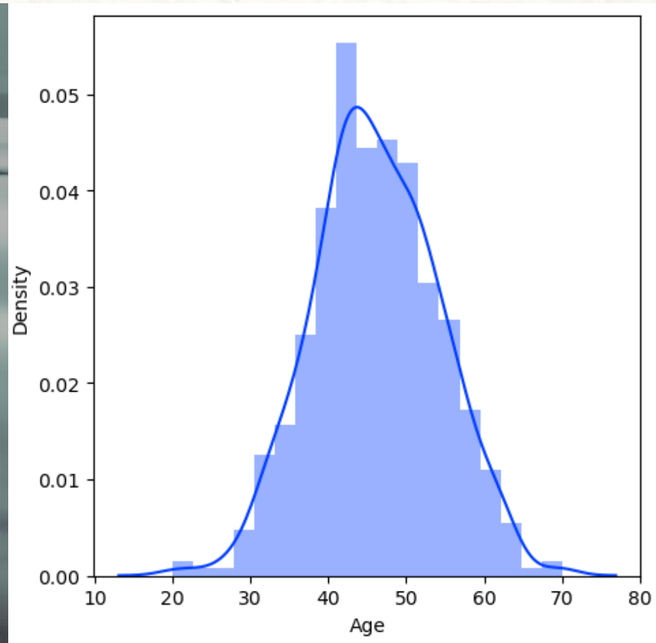

Cafelife 1890

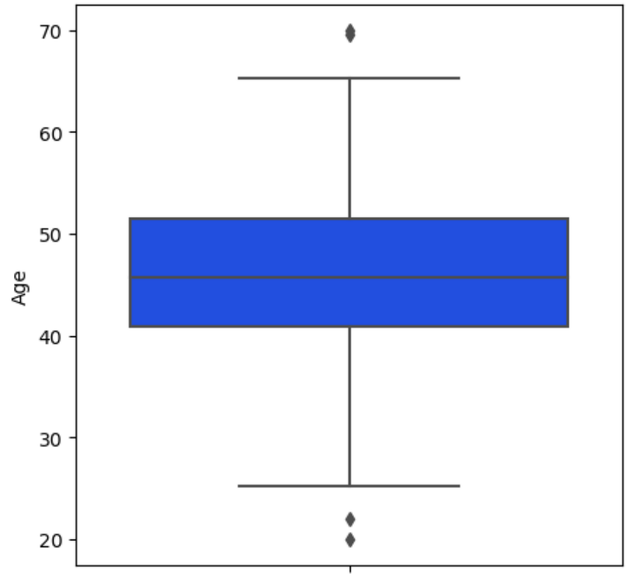

Gender
0.0
1.0
Count
Age
20
30
40
50
60
70

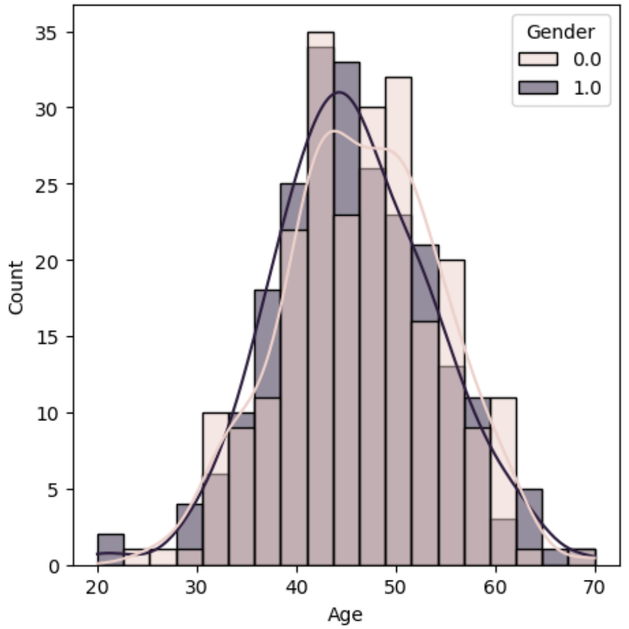


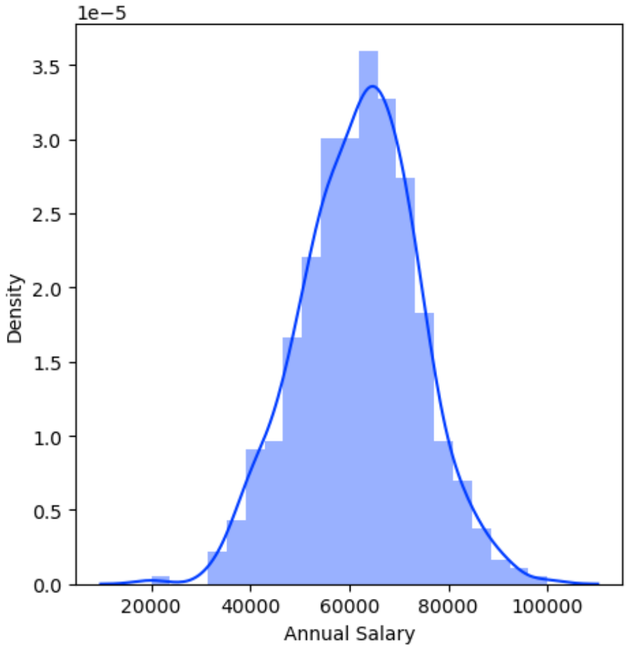

Gender
0.0
1.0
Count
Annual Salary
20000
30000
40000
50000
60000
70000
80000
90000
100000

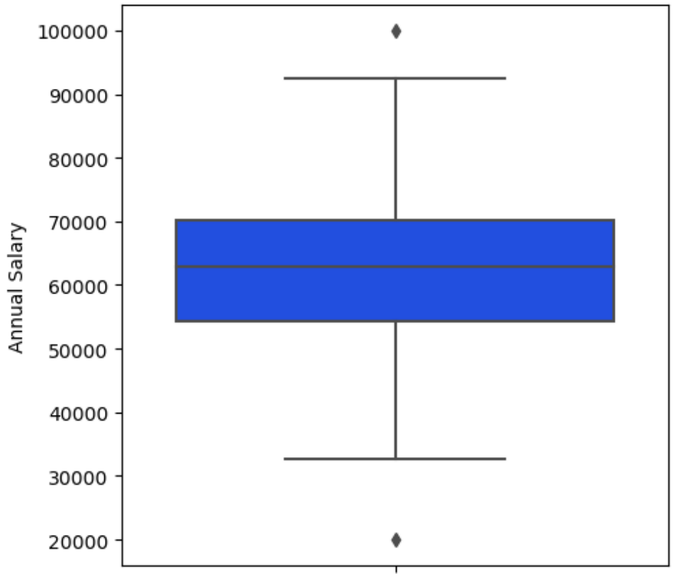
Visualizing continious columns

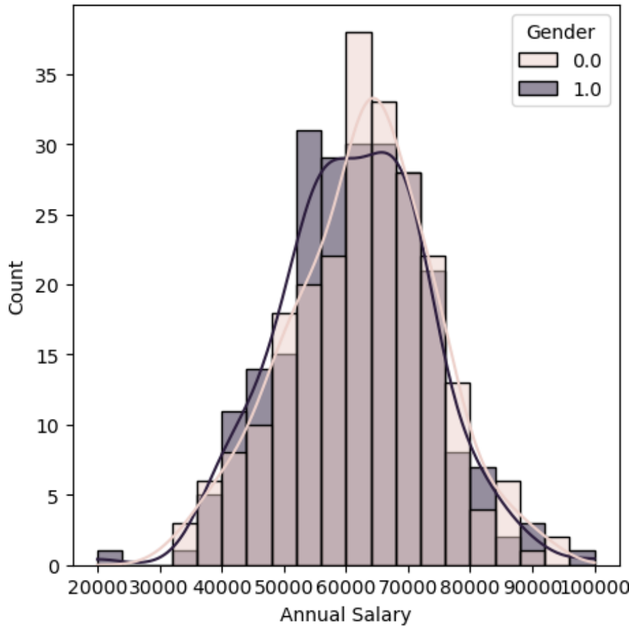




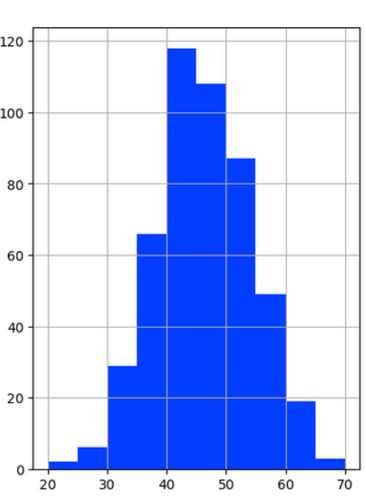


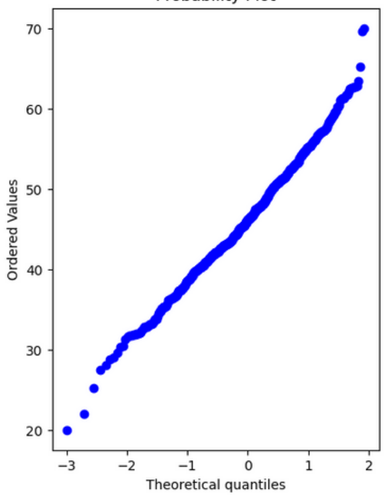


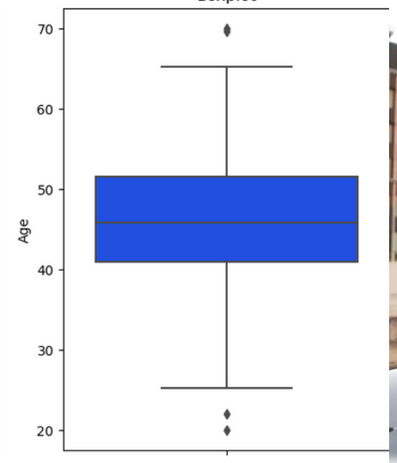


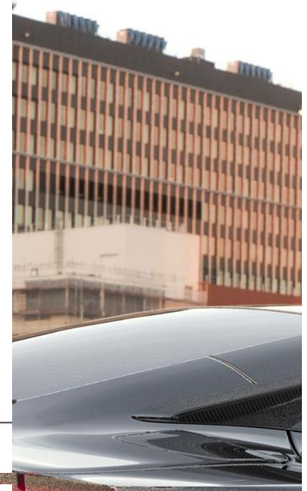


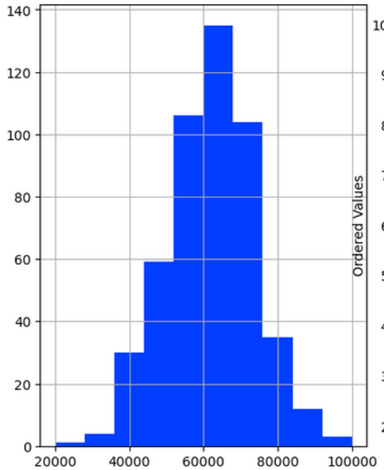
Diagnostic Plots

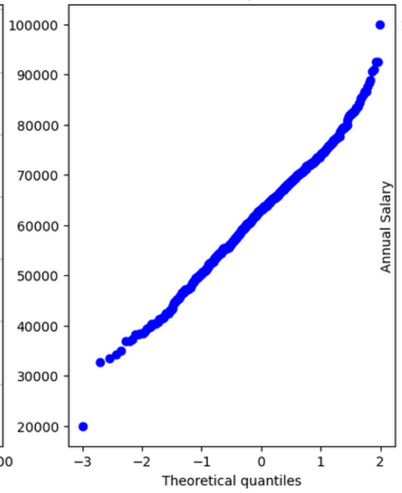


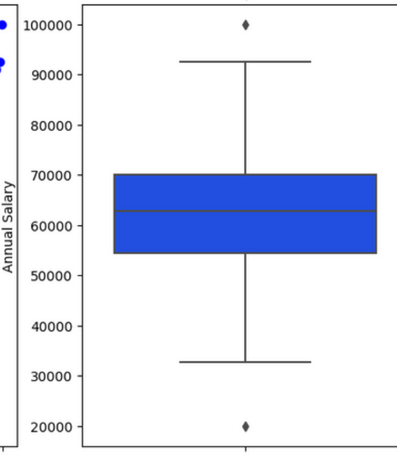


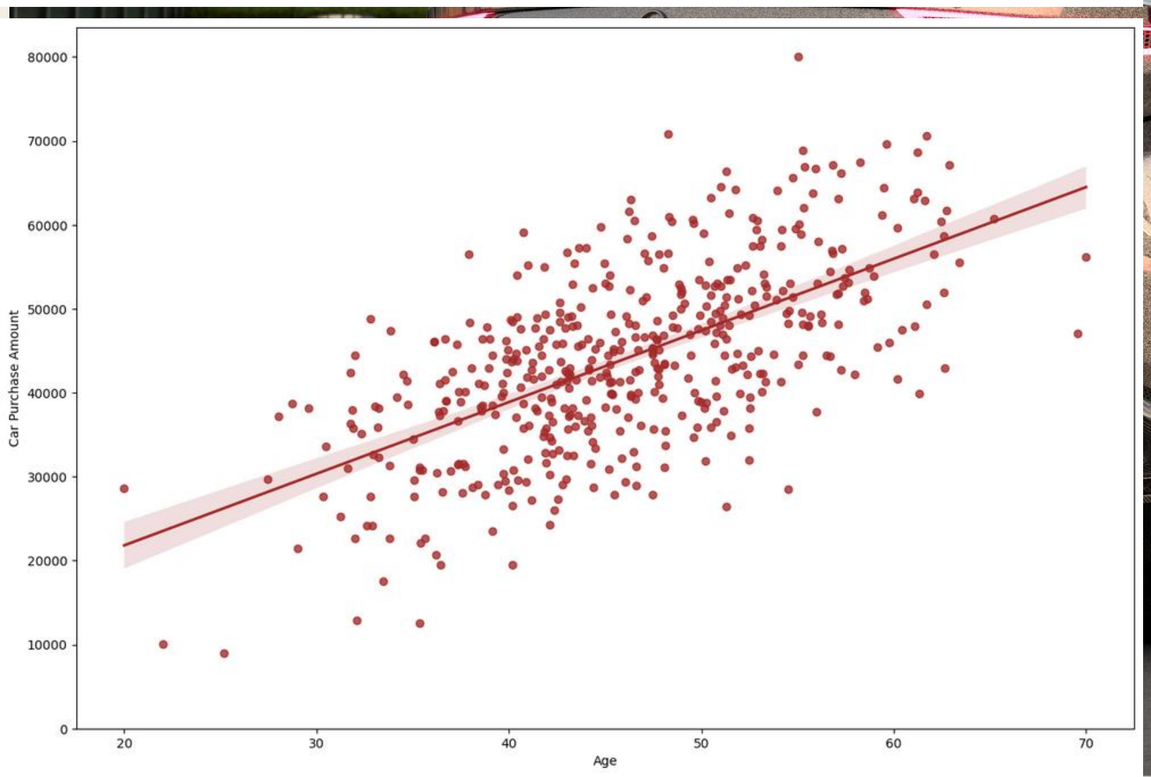


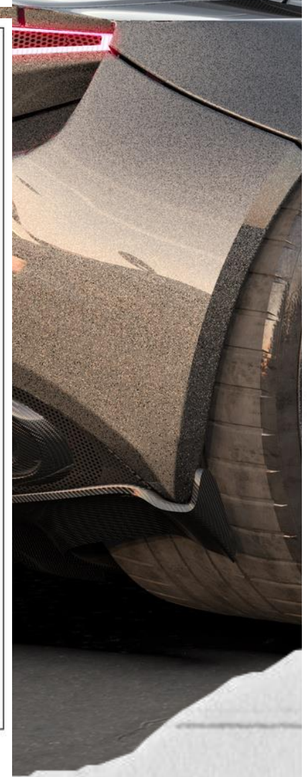


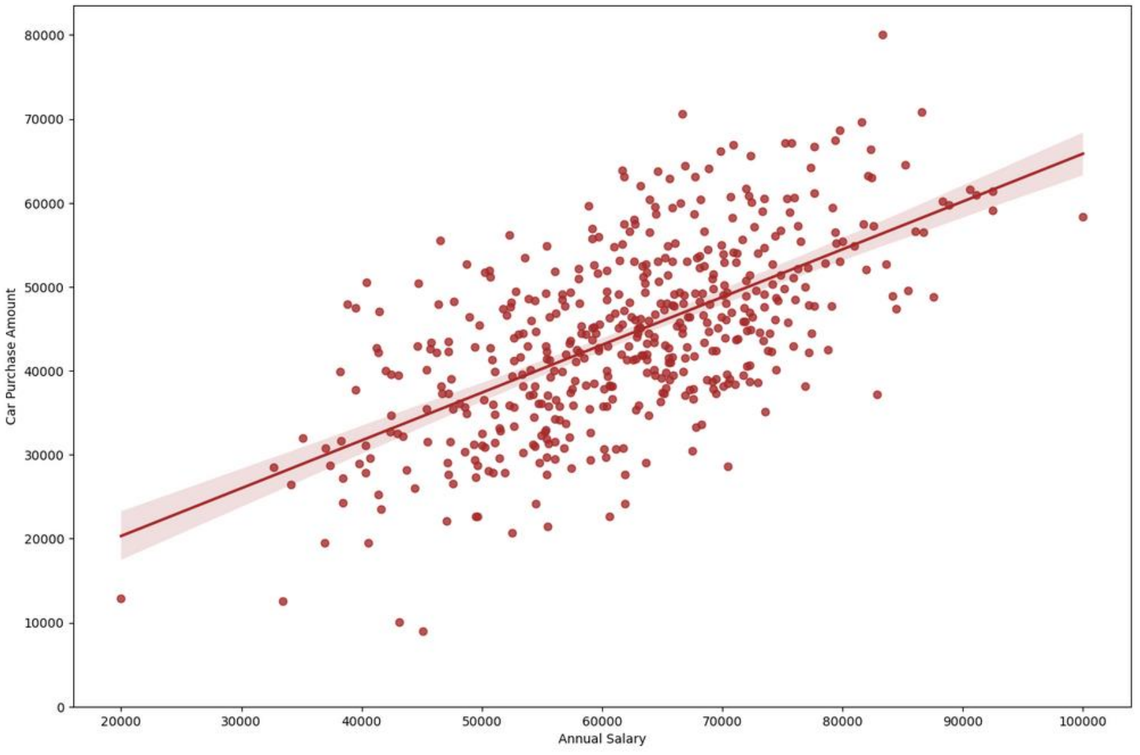


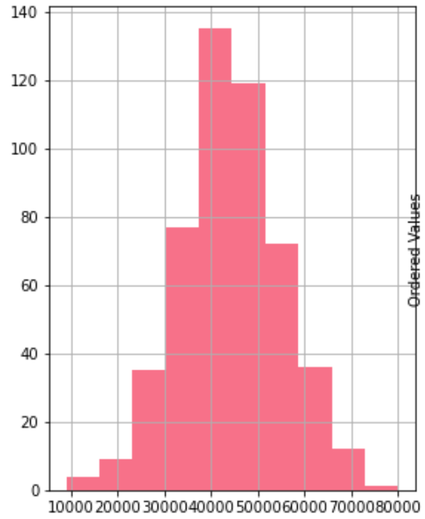


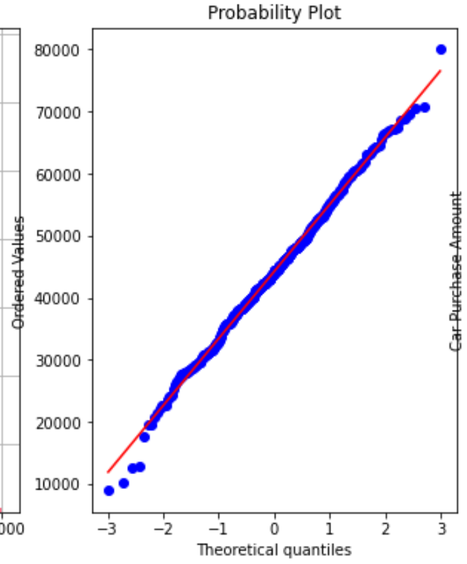


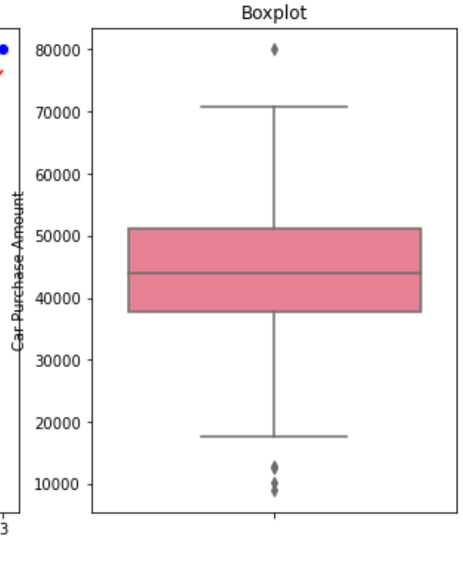


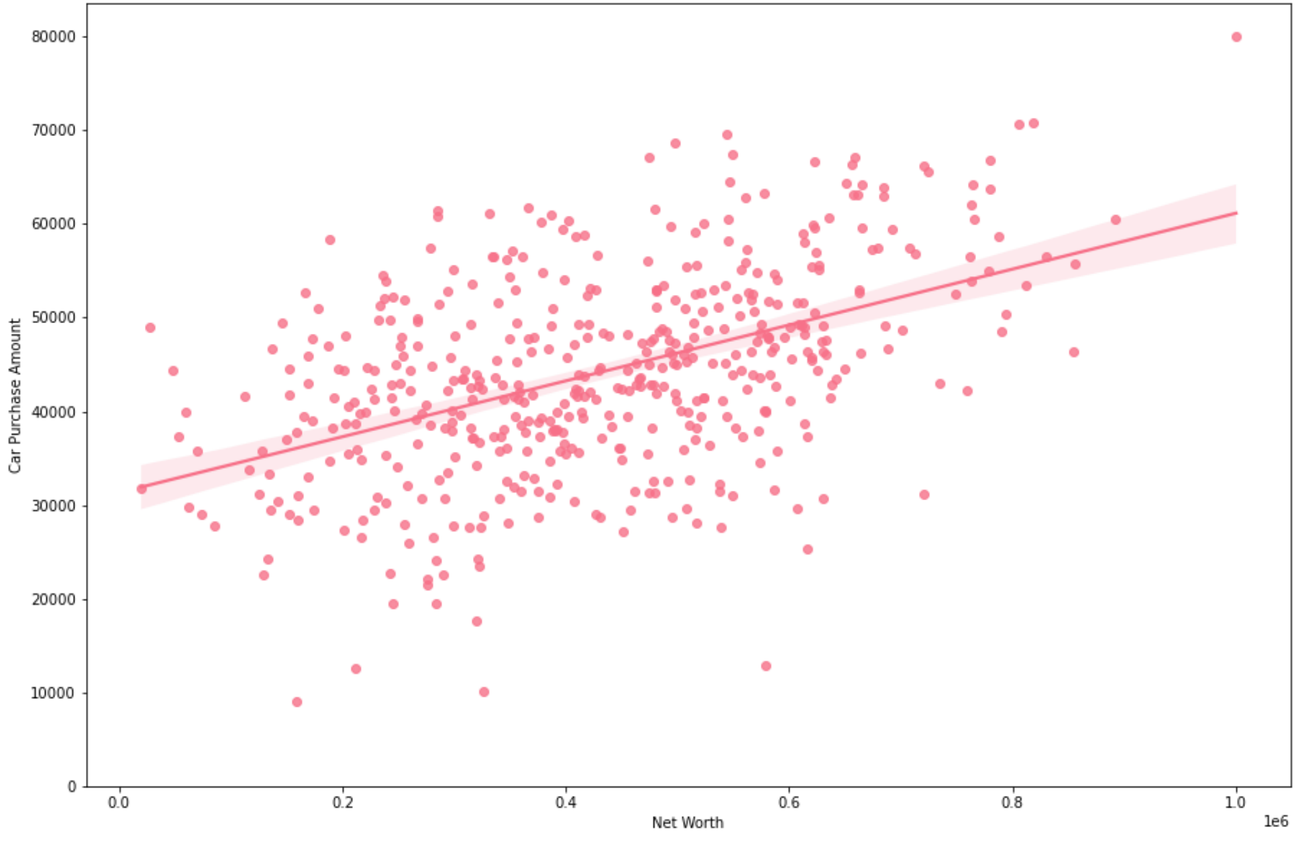





140
120
100
80
60
40
20
0
10000 20000 30000 40000 50000 60000 70000 80000


Probability Plot
Ordered Values
80000
70000
60000
50000
40000
30000
20000
10000
−3 −2 −1 0 1 2 3
Theoretical quantiles


Boxplot
Car Purchase Amount
80000
70000
60000
50000
40000
30000
20000
10000

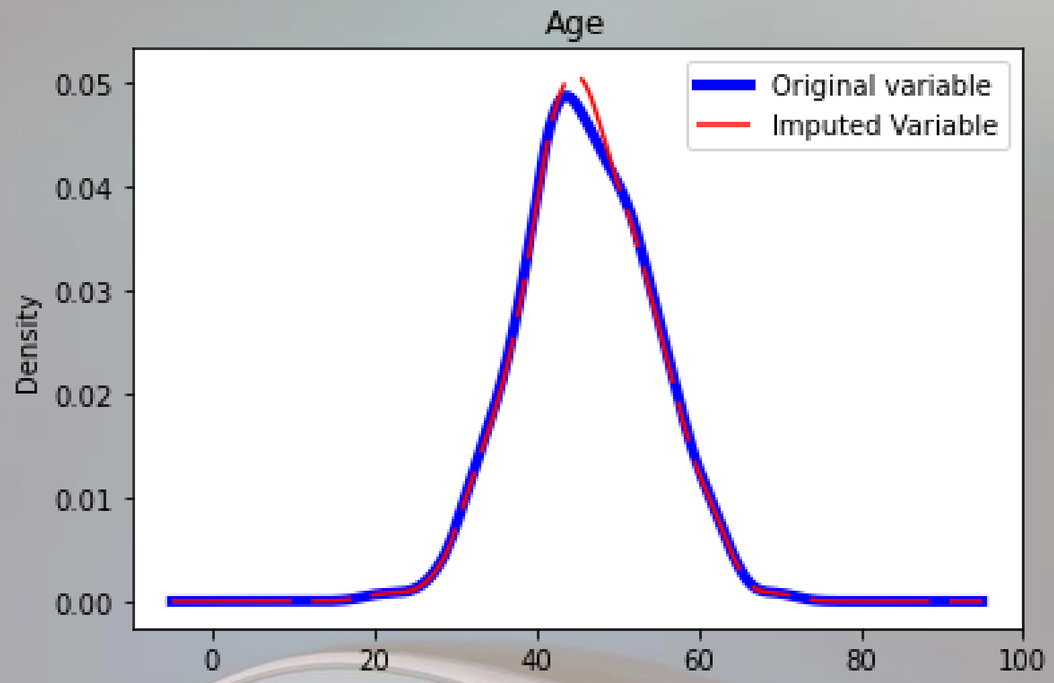

Car Purchase Amount
80000
70000
60000
50000
40000
30000
20000
10000
0
0.0 0.2 0.4 0.6 0.8 1.0
Net Worth
1e6

Comparison Age: Original vs Imputed

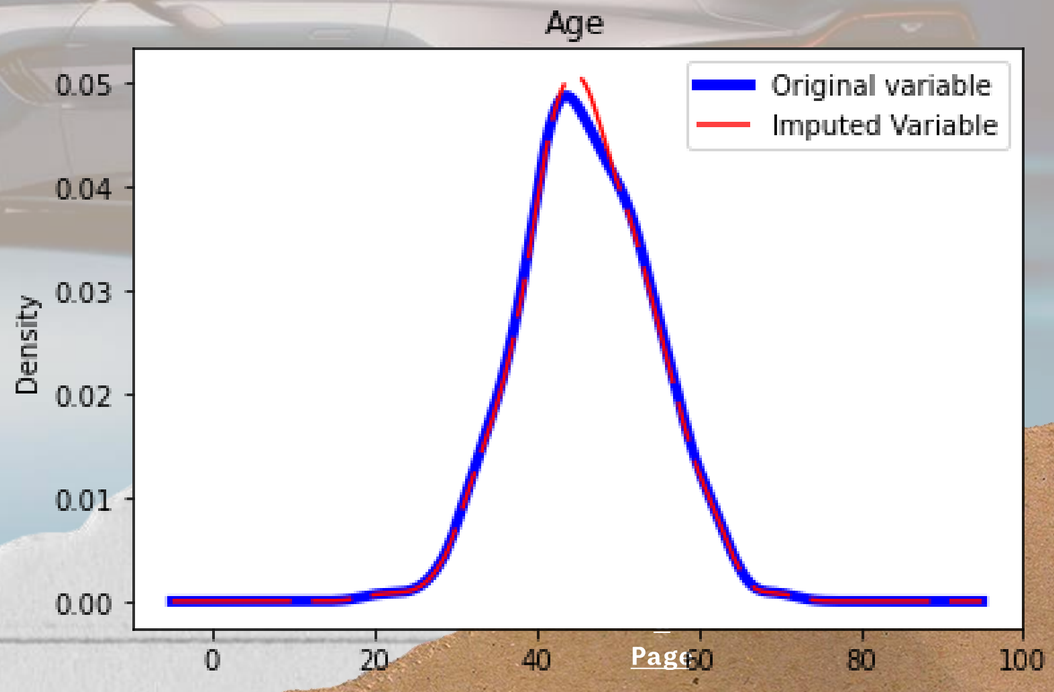
Imp. Mean

Original Variable Variance: 63.22

Variance After Imputation: 61.58



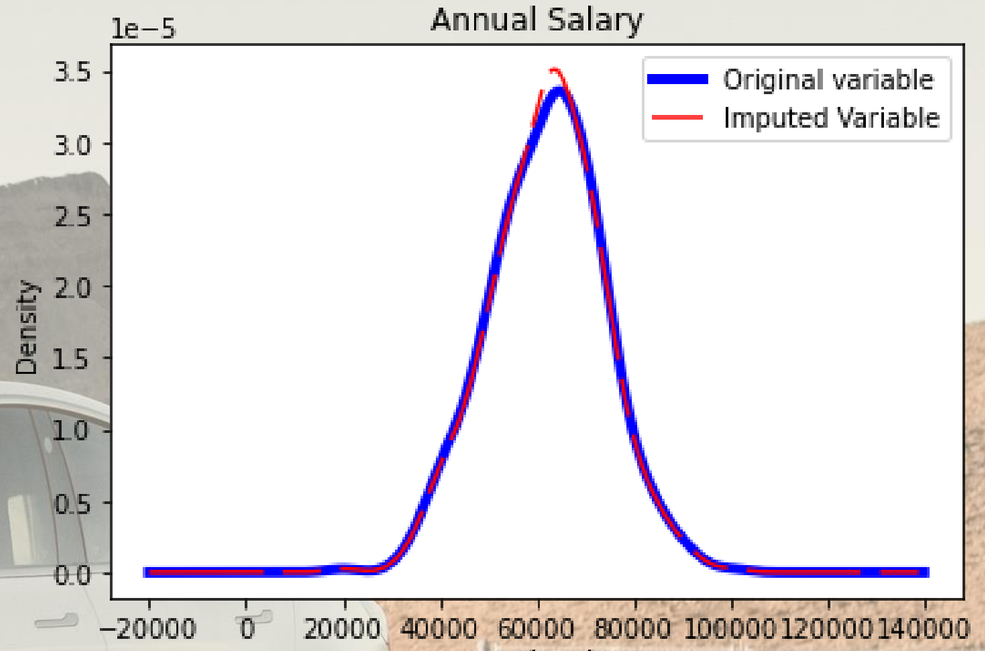
Imp. Median

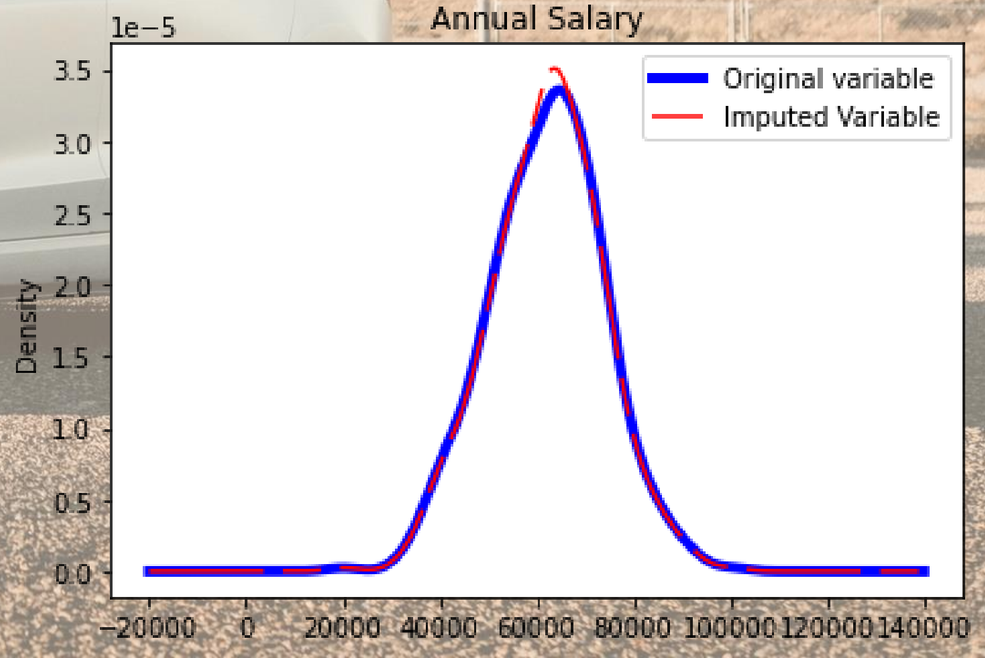


Comparison Annual Salary: Original vs Imputed

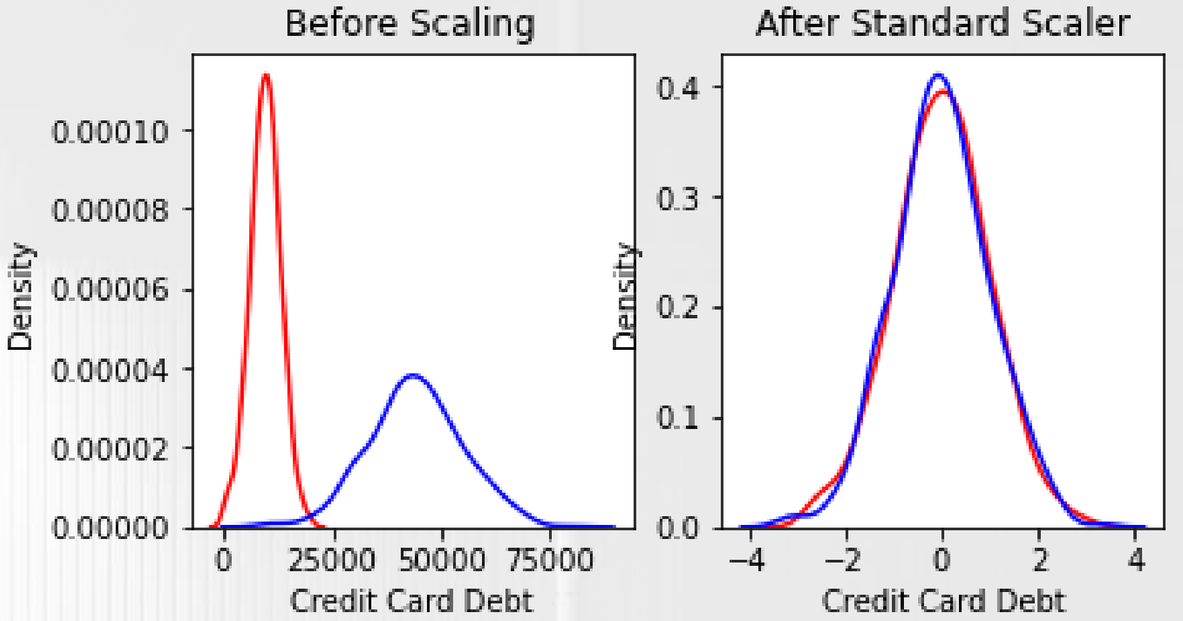
Imp. Mean

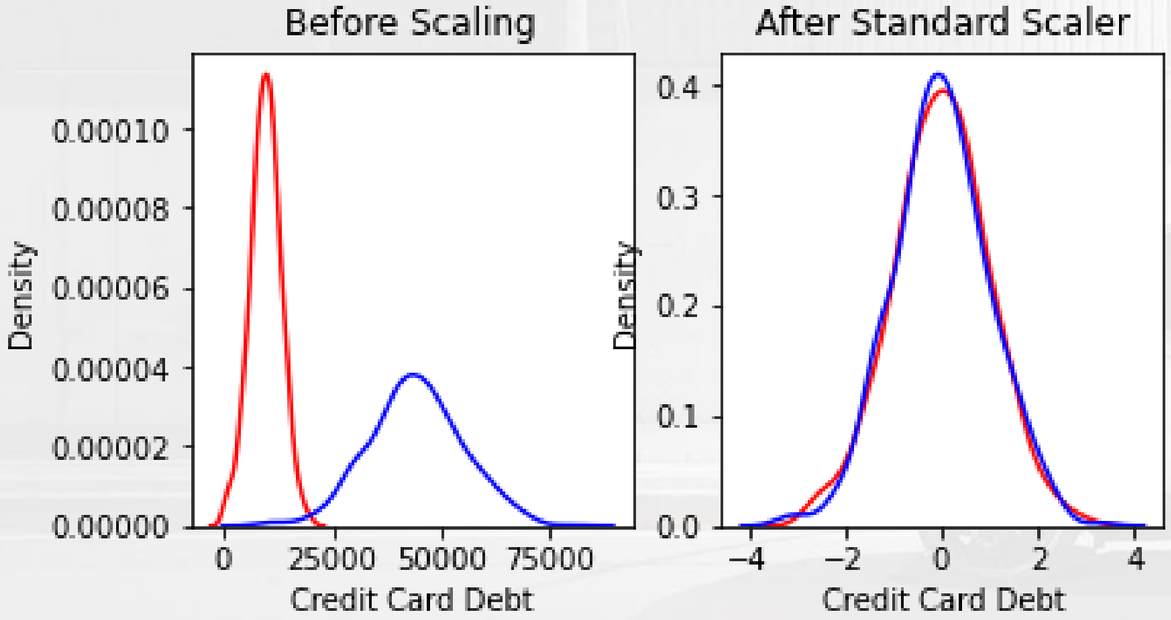
Imp. Median

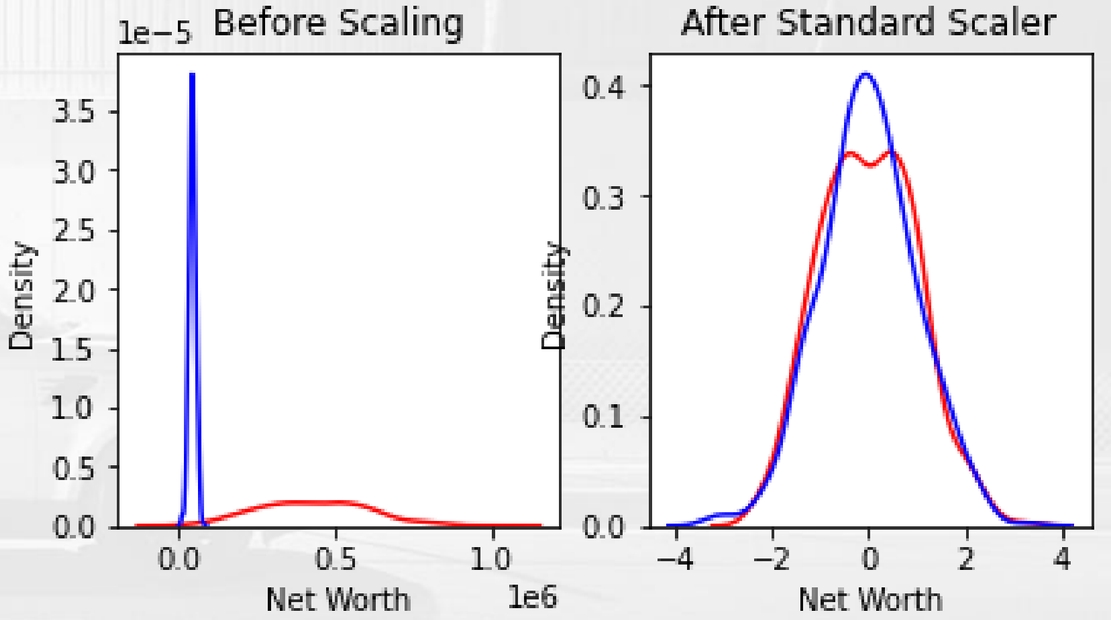




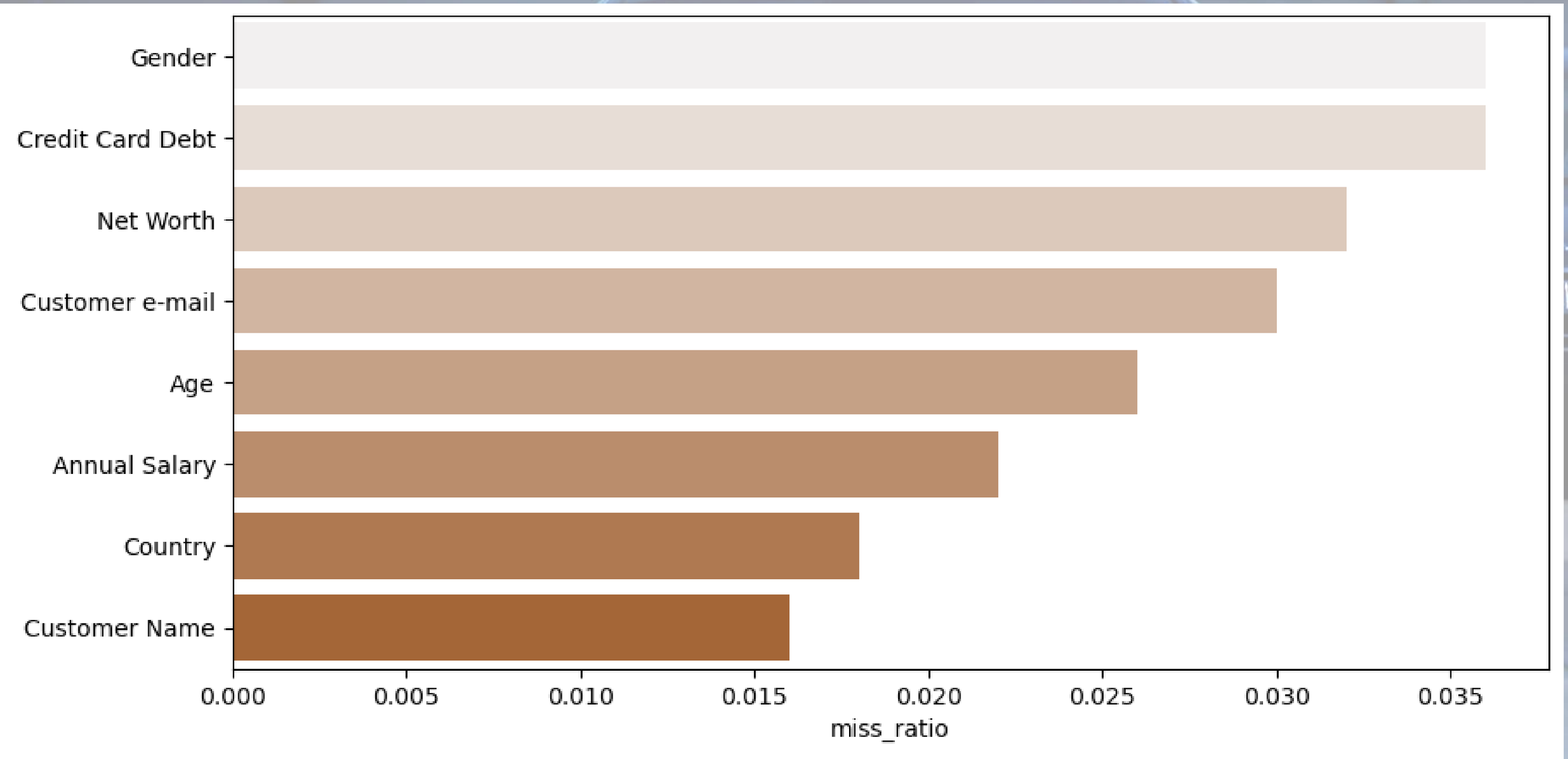
Comparison before and after scaling

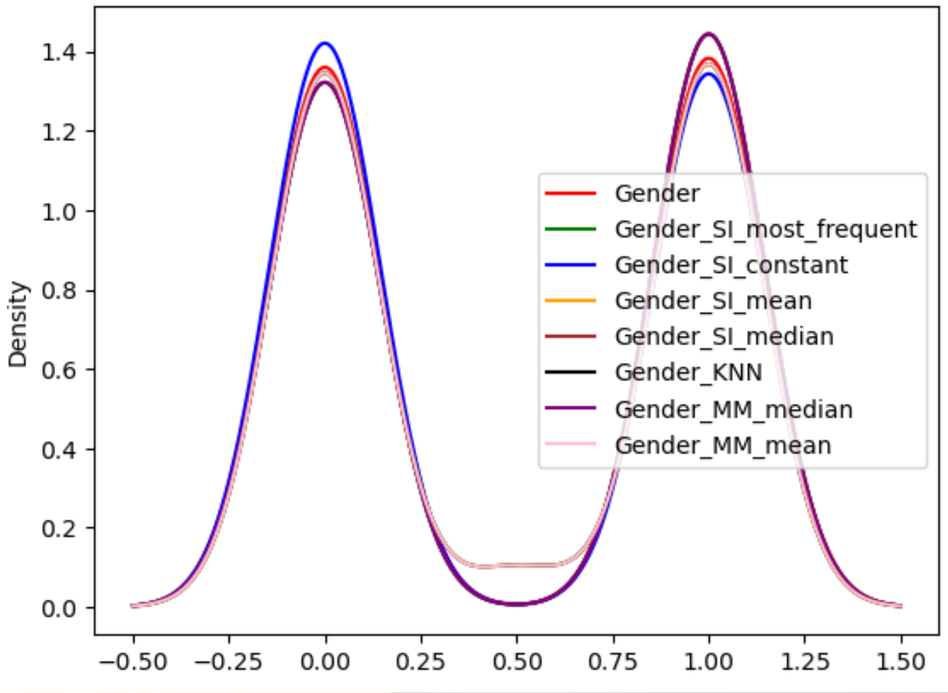


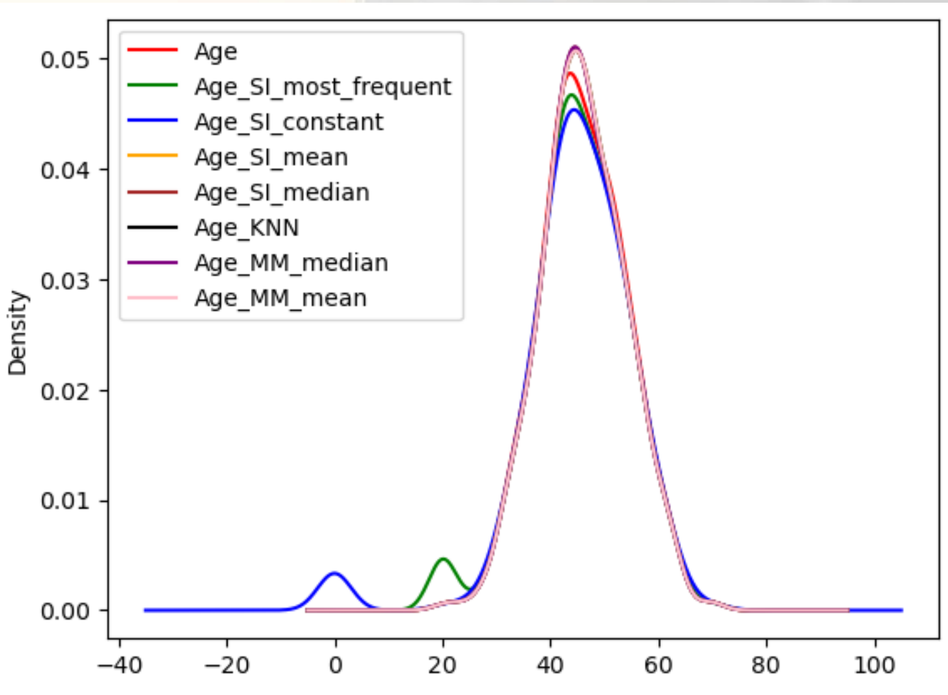


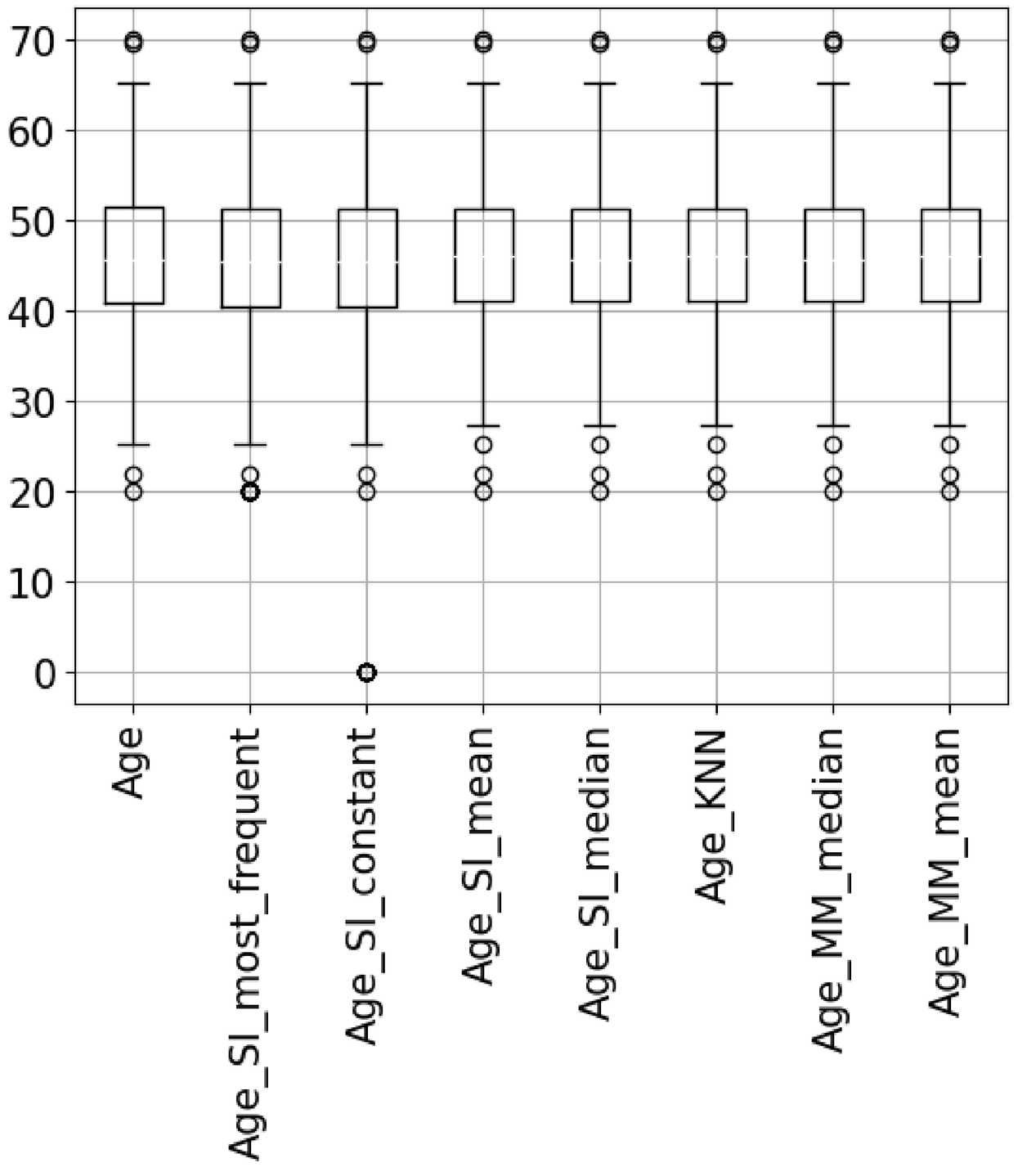


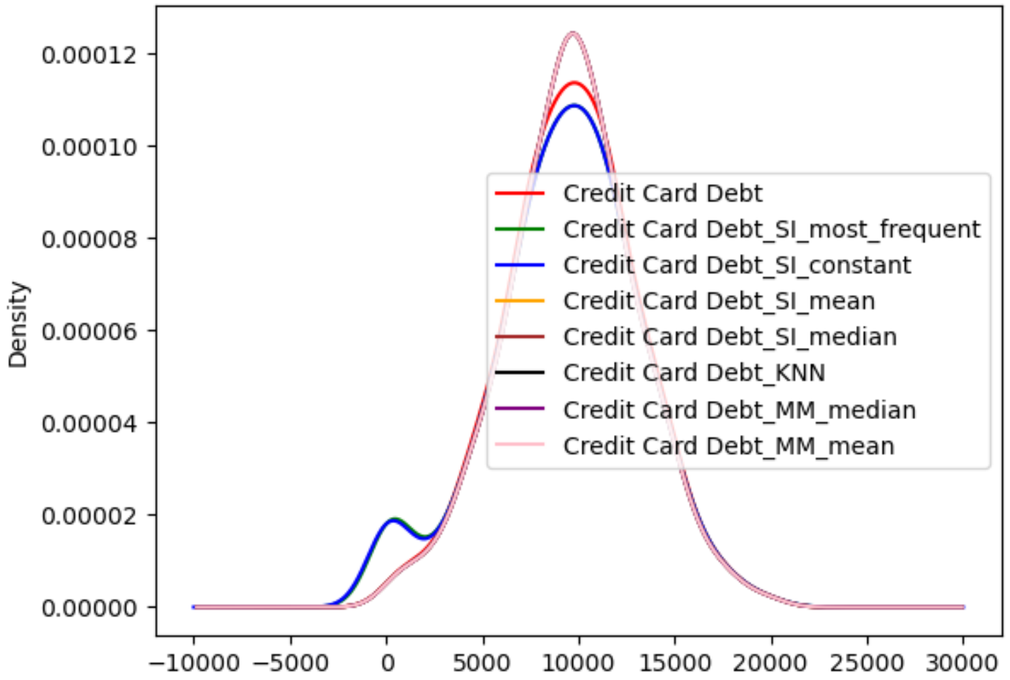
Missing values ratio

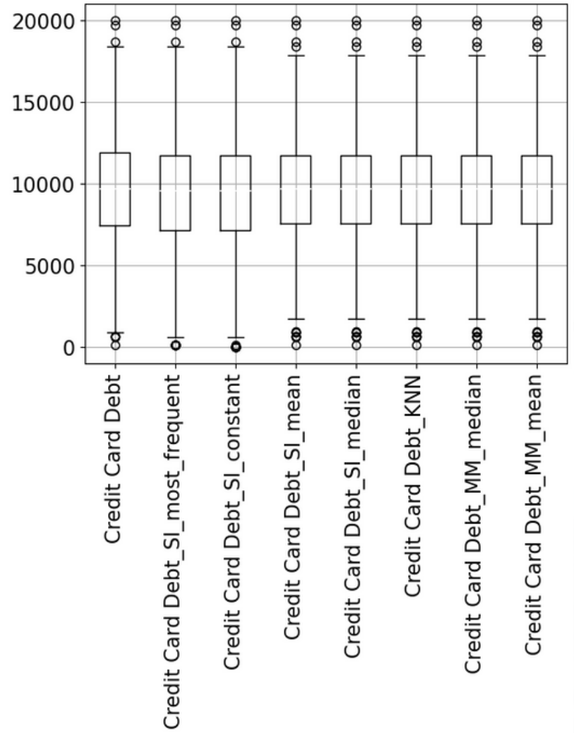


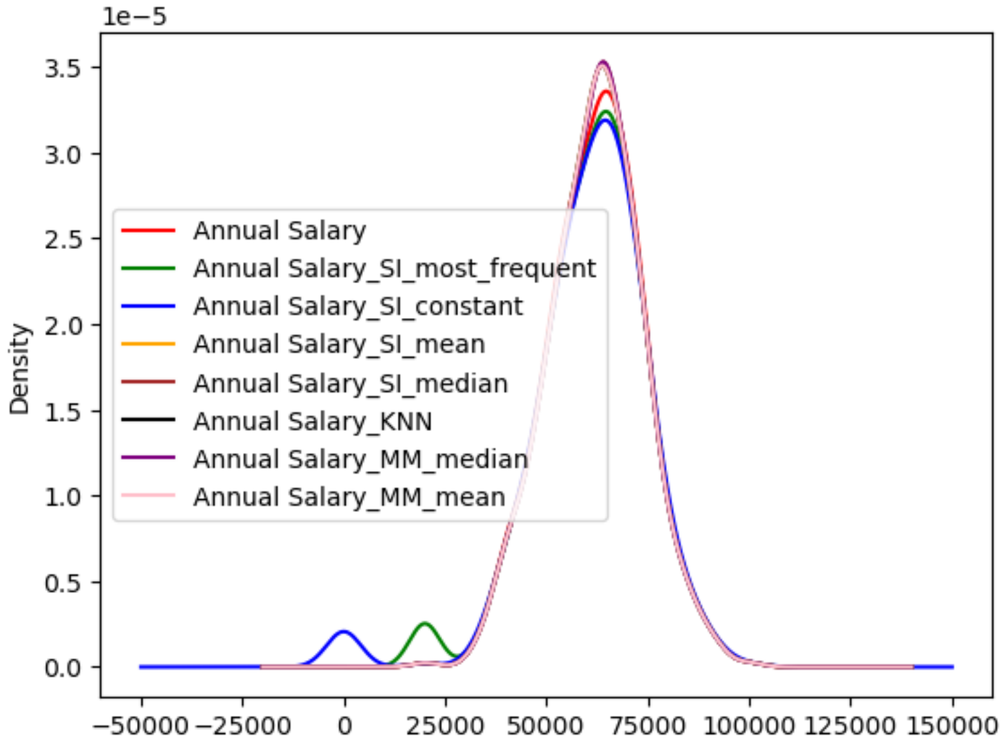

Density
Gender
Gender_SI_most_frequent
Gender_SI_constant
Gender_SI_mean
Gender_SI_median
Gender_KNN
Gender_MM_median
Gender_MM_mean

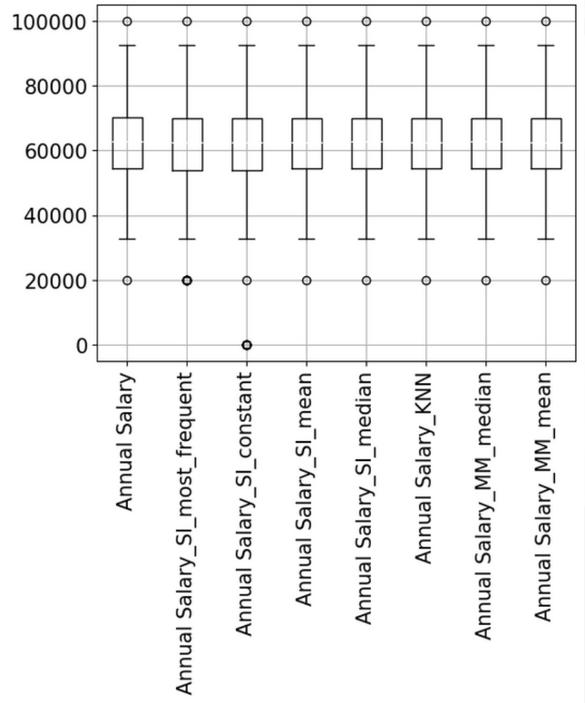

Density
Age
Age_SI_most_frequent
Age_SI_constant
Age_SI_mean
Age_SI_median
Age_KNN
Age_MM_median
Age_MM_mean


Age
Age_SI_most_frequent
Age_SI_constant
Age_SI_mean
Age_SI_median
Age_KNN
Age_MM_median
Age_MM_mean


Density
0.00012
0.00010
0.00008
0.00006
0.00004
0.00002
0.00000
−10000 −5000 0 5000 10000 15000 20000 25000 30000
Credit Card Debt
Credit Card Debt_SI_most_frequent
Credit Card Debt_SI_constant
Credit Card Debt_SI_mean
Credit Card Debt_SI_median
Credit Card Debt_KNN
Credit Card Debt_MM_median
Credit Card Debt_MM_mean


20000
15000
10000
5000
0
Credit Card Debt
Credit Card Debt_SI_most_frequent
Credit Card Debt_SI_constant
Credit Card Debt_SI_mean
Credit Card Debt_SI_median
Credit Card Debt_KNN
Credit Card Debt_MM_median
Credit Card Debt_MM_mean


1e−5
Density
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
−50000 −25000 0 25000 50000 75000 100000 125000 150000
Annual Salary
Annual Salary_SI_most_frequent
Annual Salary_SI_constant
Annual Salary_SI_mean
Annual Salary_SI_median
Annual Salary_KNN
Annual Salary_MM_median
Annual Salary_MM_mean

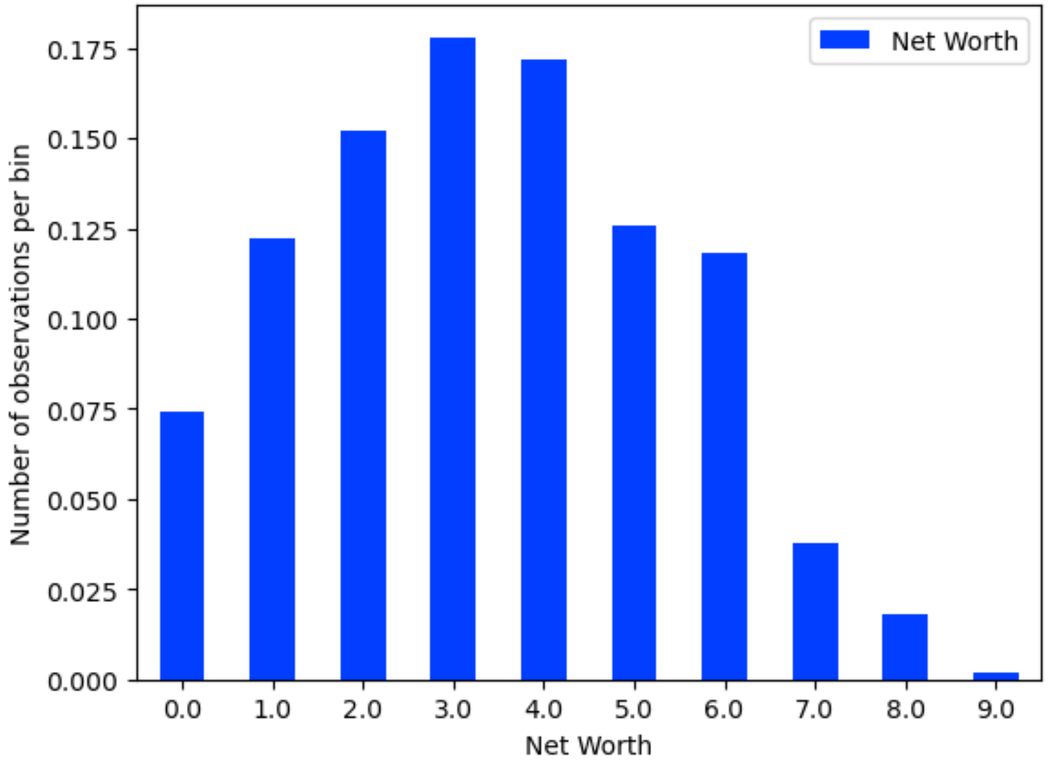

100000
80000
60000
40000
20000
0
Annual Salary
Annual Salary_SI_most_frequent
Annual Salary_SI_constant
Annual Salary_SI_mean
Annual Salary_SI_median
Annual Salary_KNN
Annual Salary_MM_median
Annual Salary_MM_mean

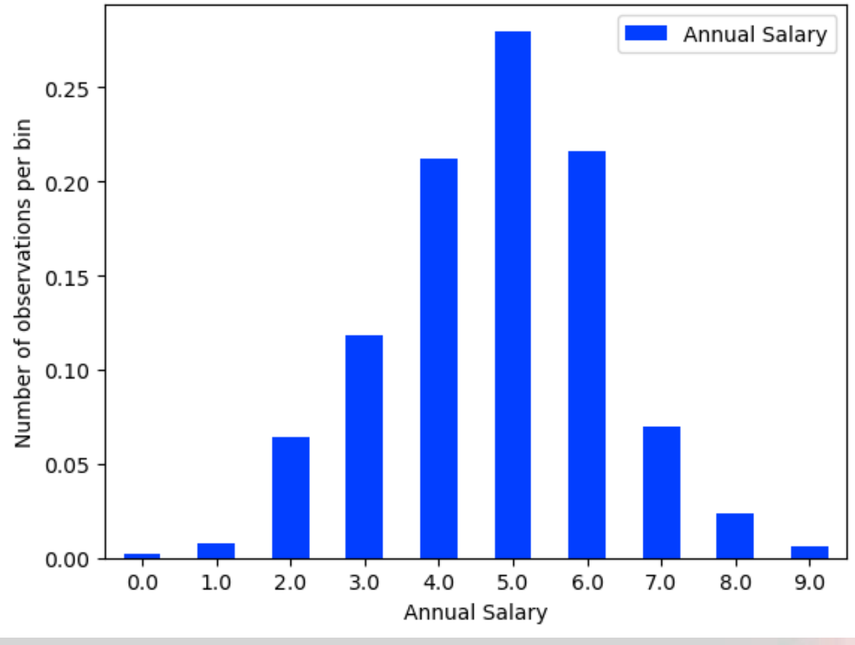

長野 95
せ 2JZGE

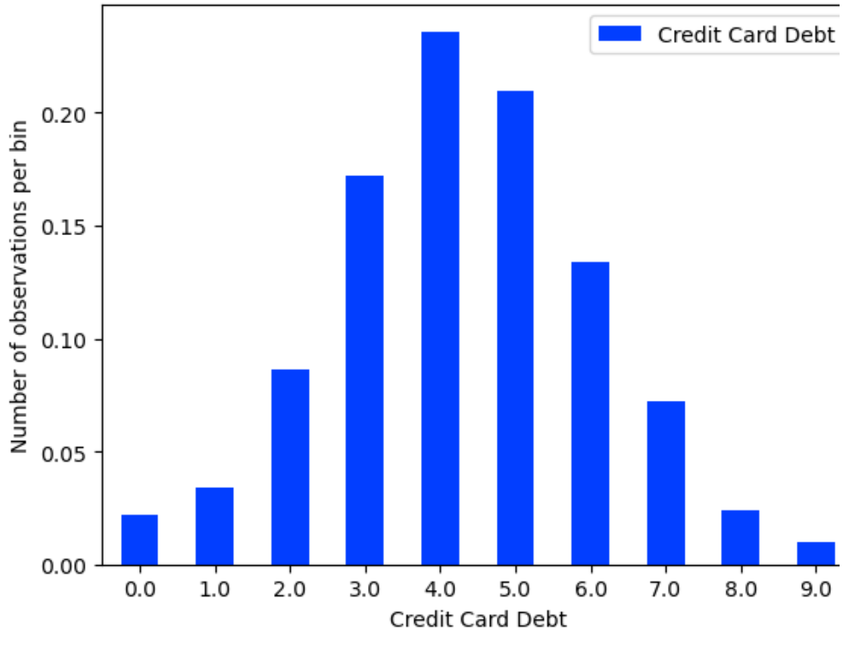

1e6
1.0
0.8
0.6
0.4
0.2
0.0
Net Worth
Net Worth_SI_most_frequent
Net Worth_SI_constant
Net Worth_SI_mean
Net Worth_SI_median
Net Worth_KNN
Net Worth_MM_median
Net Worth_MM_mean
1e−6
2.0
1.5
1.0
0.5
0.0
Density
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1e6
Net Worth
Net Worth_SI_most_frequent
Net Worth_SI_constant
Net Worth_SI_mean
Net Worth_SI_median
Net Worth_KNN
Net Worth_MM_median
Net Worth_MM_mean

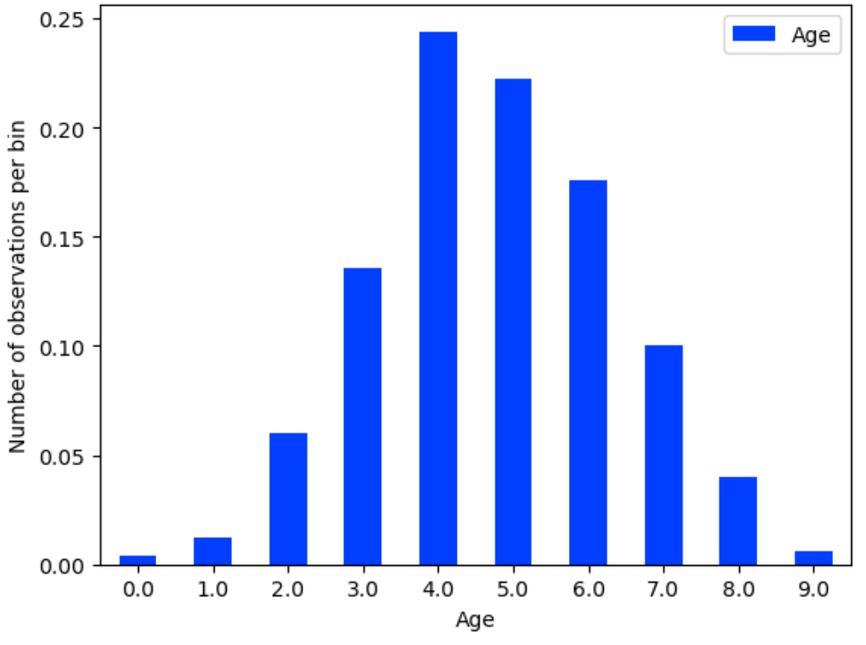
Discretisation method

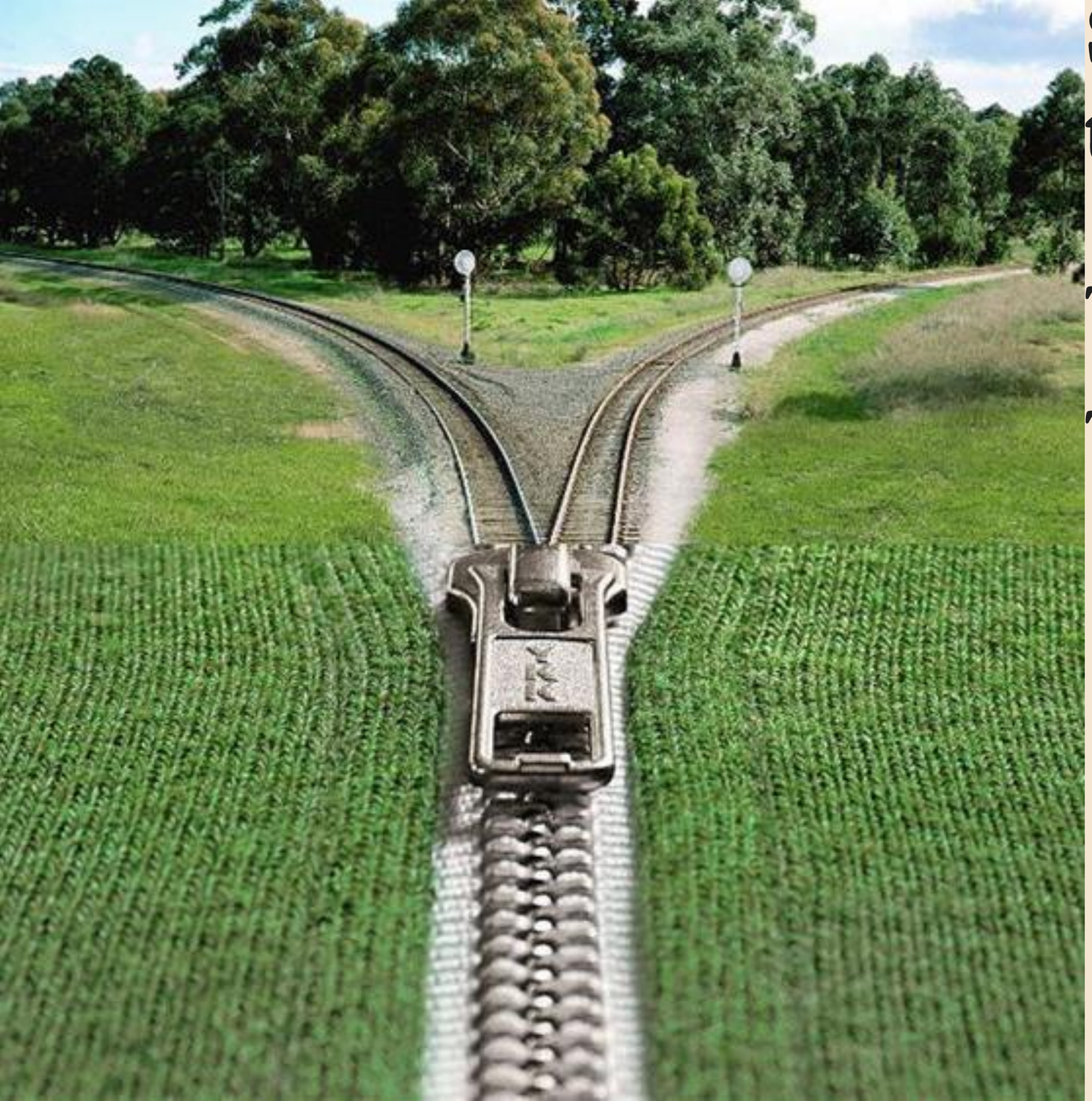
KBinsDiscretizer











Splitting dataset into training and test sets

Training set: (350, 5)

Test set: (150, 5)

RandomForestRegressor

Model training

Best parameters GridSearchCV:

RFR__criterion : squared_error ,
RFR__max_depth : 3,
RFR__min_samples_leaf : 1,
RFR__min_weight_fraction_leaf : 0.0,
RFR__n_estimators : 10,
discretizer__encode : onehot ,
discretizer__n_bins : 5,
discretizer__strategy : uniform ,
imputer imputation method :

Metrics scores:

train mse:
5075.480132750794
train rmse:
42578294.11384401
train r2:
0.638396664513572 6

test mse:
6290.812176432834
test rmse:
65528615.06108584
test r2:
0.410230305156430 8

K-Nearest Neighbor Regression

Best parameters GridSearchCV:

```
KNN__metric : minkowski ,
KNN__n_neighbors : 5,
KNN__p : 6,
discretizer__encode :
ordinal ,
discretizer__n_bins : 10,
discretizer__strategy :
quantile ,
imputer__imputation_method :
median
```

Metrics scores:

```
train mse: 2328.6032296171425
train rmse: 9952988.24222657
train r2: 0.9154725706759557

test mse: 3043.2516187066667
test rmse: 18422981.34763793
test r2: 0.834189749357948
```

Linear Regression

Best parameters GridSearchCV:

```
LR__fit_intercept : True,
discretizer__encode :
ordinal ,
discretizer__n_bins : 10,
discretizer__strategy :
kmeans ,
imputer__imputation_method :
mean
```

Metrics scores:

```
train mse:  1659.347184287053
train rmse: 5014026.472393233
train r2:   0.957417535522046

test mse:  2161.022499500849
test rmse: 11076037.68760471
test r2:   0.900313605751357
```

Support Vector Regression

Best parameters GridSearchCV:

```
SVR__C : 150.0,
SVR__degree : 1,
SVR__epsilon : 2.0,
SVR__kernel : linear ,
discretizer__encode :
ordinal ,
discretizer__n_bins : 10,
discretizer__strategy :
kmeans ,
imputer__imputation_method :
mean
```

Metrics scores:

```
train mse:  1718.3957797495216
train rmse: 5490166.1223816415
train r2:   0.9533738393341997

test mse:  2166.392792217919
test rmse: 11059244.182386532
test r2:   0.900464750238251 9
```

XGBoost Regression

Best parameters GridSearchCV:

```
XGB__max_depth : 1,
XGB__max_leaves : 0,
XGB__n_estimators : 500,
discretizer__encode :
onehot-dense ,
discretizer__n_bins : 10,
discretizer__strategy :
kmeans ,
imputer__imputation_method :
mean
```

Metrics scores:

```
train mse:  1520.4276128116073
train rmse: 4426711.764099453
train r2:   0.962405404621062 3

test mse:  2194.6212560416666
test rmse: 11274341.715132743
test r2:   0.898528832530681 3
```

AdaBoost Regressor

Best parameters GridSearchCV:

```
AB__learning_rate : 1.5,
AB__loss : square ,
AB__n_estimators : 300,
discretizer__encode :
ordinal ,
discretizer__n_bins : 10,
discretizer__strategy :
kmeans ,
imputer__imputation_method :
mean
```

Metrics scores:

```
train mse:  2841.1505059426895
train rmse: 12194798.540033158
train r2:   0.896433619067250 7

test mse:  3639.5861699107754
test rmse: 24782993.297038626
test r2:   0.776948461668548 9
```

Gradient Boosting Regression

Best parameters GridSearchCV:

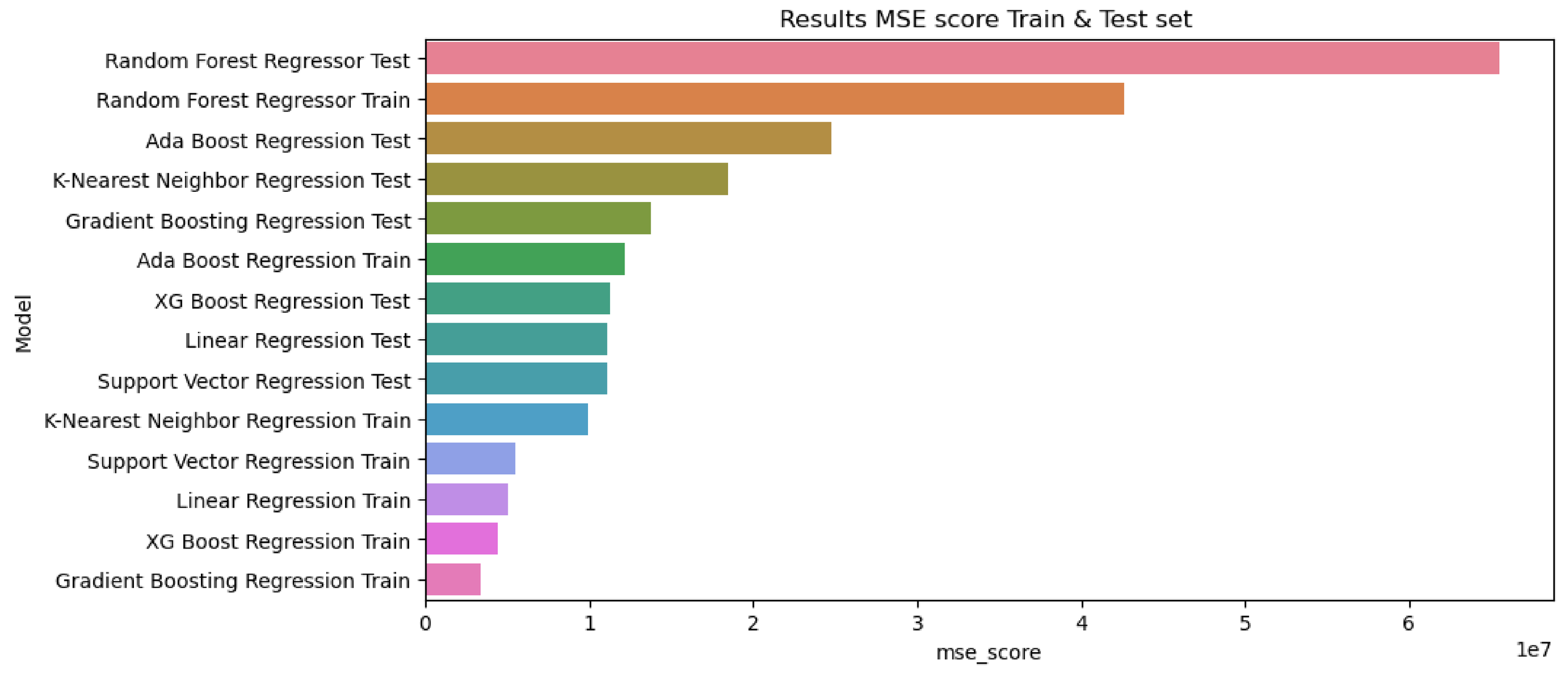
```
GB__learning_rate : 0.2,
GB__loss : huber ,
GB__n_estimators : 100,
discretizer__encode :
ordinal ,
discretizer__n_bins : 10,
discretizer__strategy :
kmeans ,
imputer__imputation_method :
mean
```

Metrics scores:

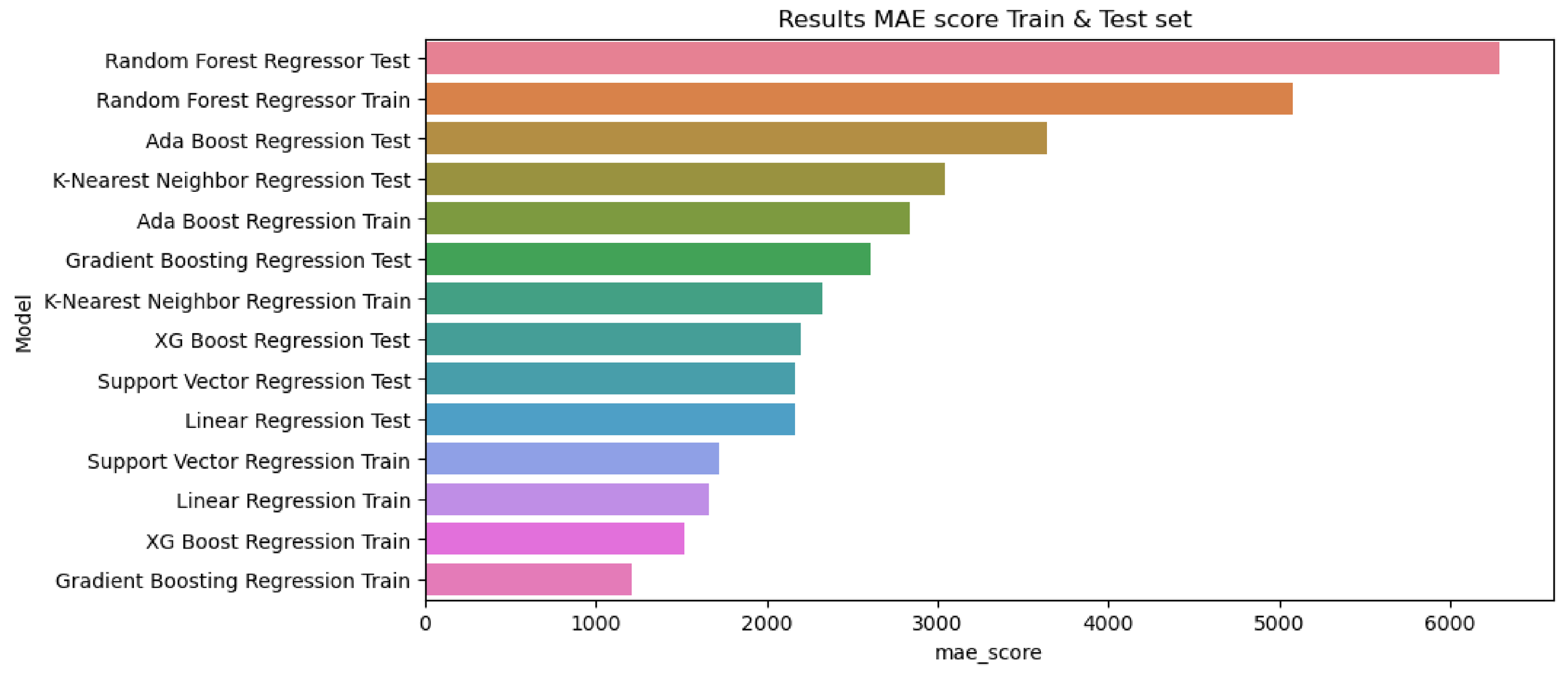
```
train mse:  1210.8664932003755
train rmse: 3351310.7544226693
train r2:   0.9715384288574223

test mse:  2604.226042804091
test rmse: 13752036.441054996
test r2:   0.8762291202437591
```

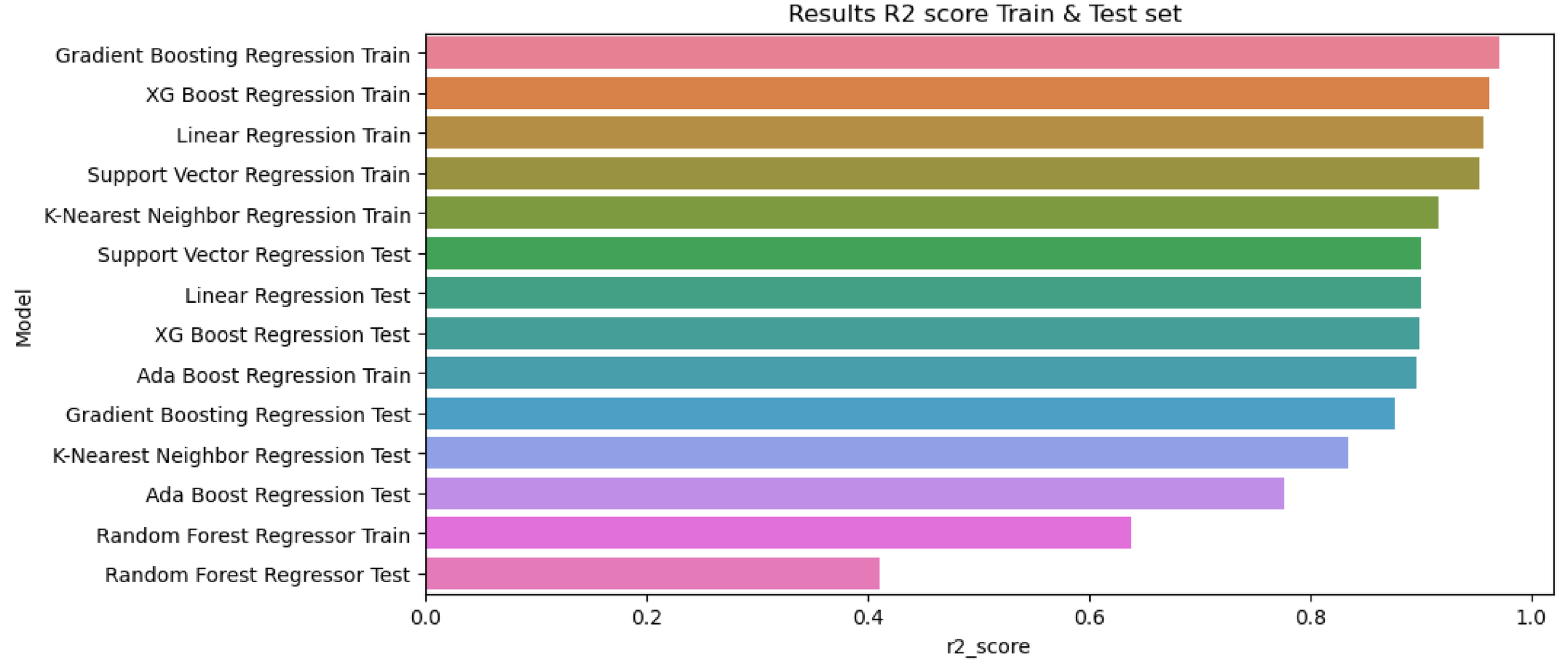
Model Evaluation



Model Evaluation



Model Evaluation



ANN

Model Structure

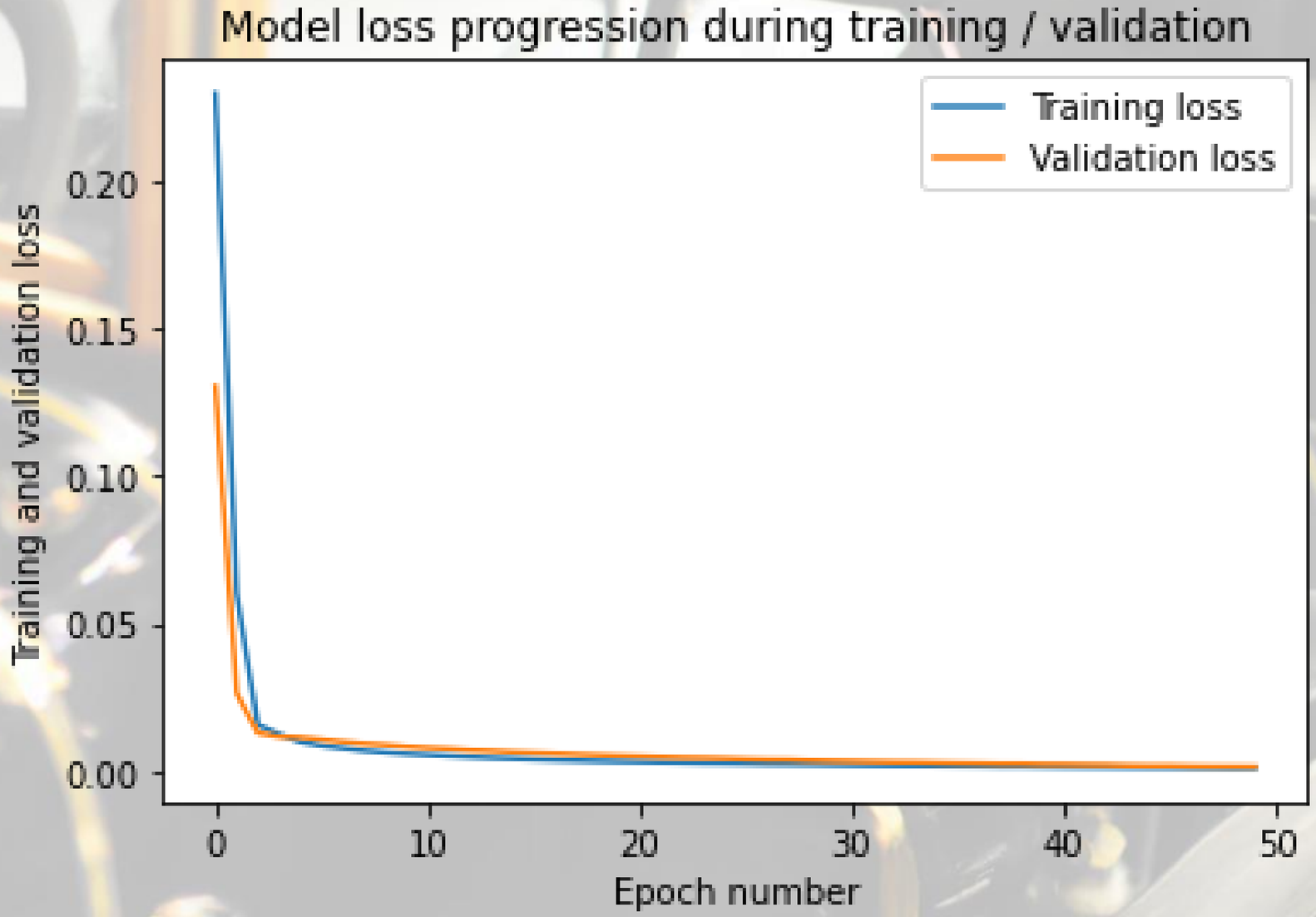
```
Model: "sequential_9"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 dense_27 (Dense)            (None, 25)                150

 dense_28 (Dense)            (None, 25)                650

 dense_29 (Dense)            (None, 1)                 26

=================================================================
Total params: 826
Trainable params: 826
Non-trainable params: 0
```

Epochs history



Model Evaluation

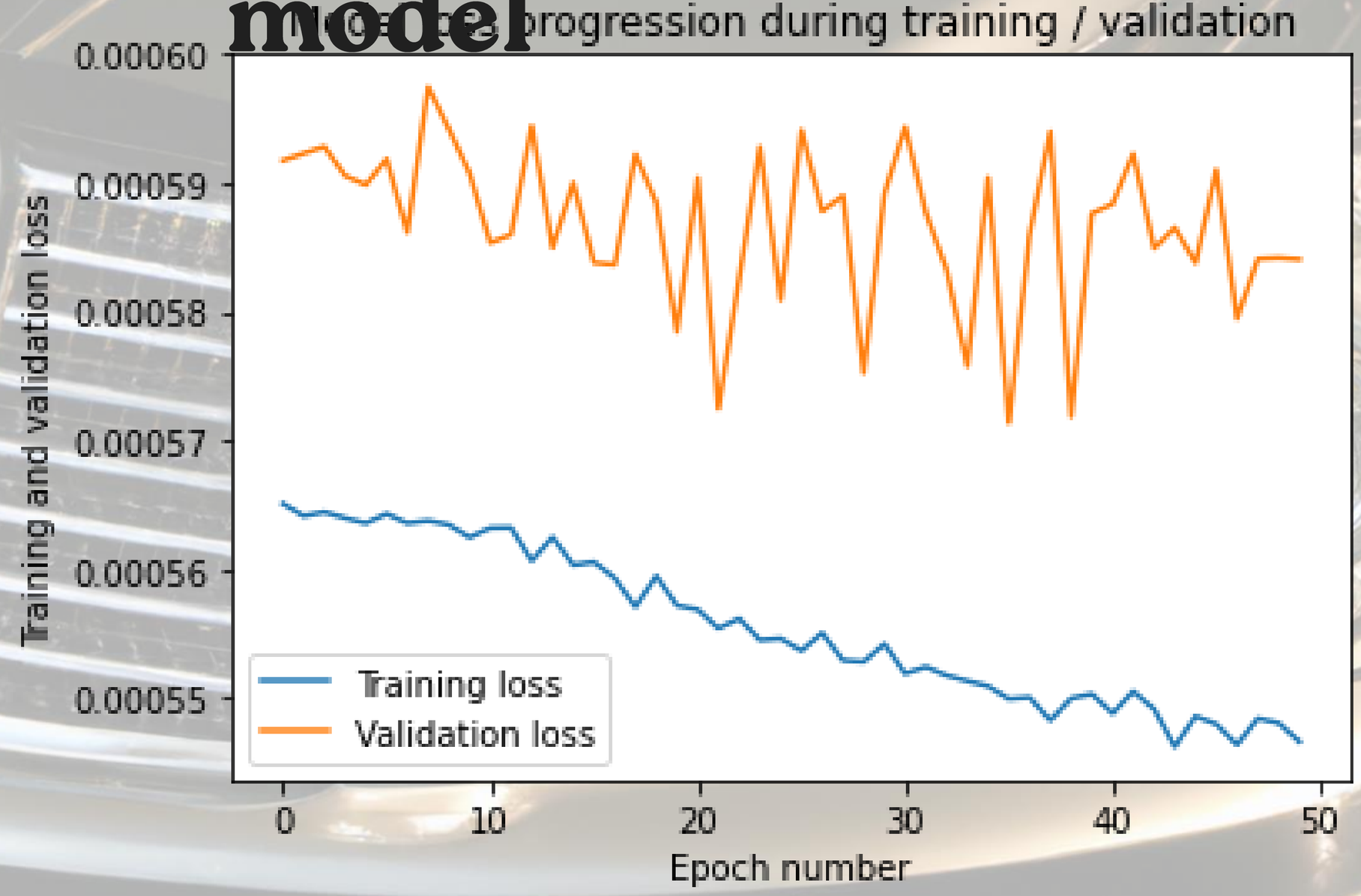
train mse: 0.033308393382620

train r2: 0.9502837976547399

test mse: 3980.137295329587

test r2: 0.8895131576934218

Additional model



Conclusion

Study Summary

I

Our goal with the machine learning phase was that we wanted to train a model which will predict car purchase amount based on certain factors that include 'Customer Name', 'Customer email', 'Country', 'Gender', 'Age', 'Annual Salary', 'Credit Card Debt' and 'Net Worth'. After the data analysis part we removed redundant variables such as 'Customer Name', 'Customer email and 'Country'. Prior to model training parts we checked how data would change if we use different imputation, scaling and discretization methods.

II

We trained eight models, one RandomForestRegressor, one KNeighborsRegressor, one LinearRegression, one SVR, one XGBRegressor, one AdaBoostRegressor, one GradientBoostingRegressor and one DecisionTreeRegressor. We also implied Pipeline and used Grid Search to find the best hyperparameters and evaluated results with the following regression metrics such as mean absolute error, mean squared error and r2 score.

III

Among ML models that we have built for this study SVR and LinearRegression showed the most stable and accurate results in both trained and test versions whereas RandomForestRegressor performed the worst among other ML models.

IV

We were somewhat successful in giving a general idea but because of limitations of time and data set we still have a long way to go. Looking forward, we would like to acquire more data for accurate prediction of price so that if someone wants to buy a used car they have an idea of what it would cost them beforehand and not end up paying more than the car’s worth.