eBay Used Car Data: Exploratory Data Analysis

Nitin Mahajan
Final Project DSC530
Bellevue University

OBJECTIVE

Identification of Significant variables to drive the price of used cars in eBay

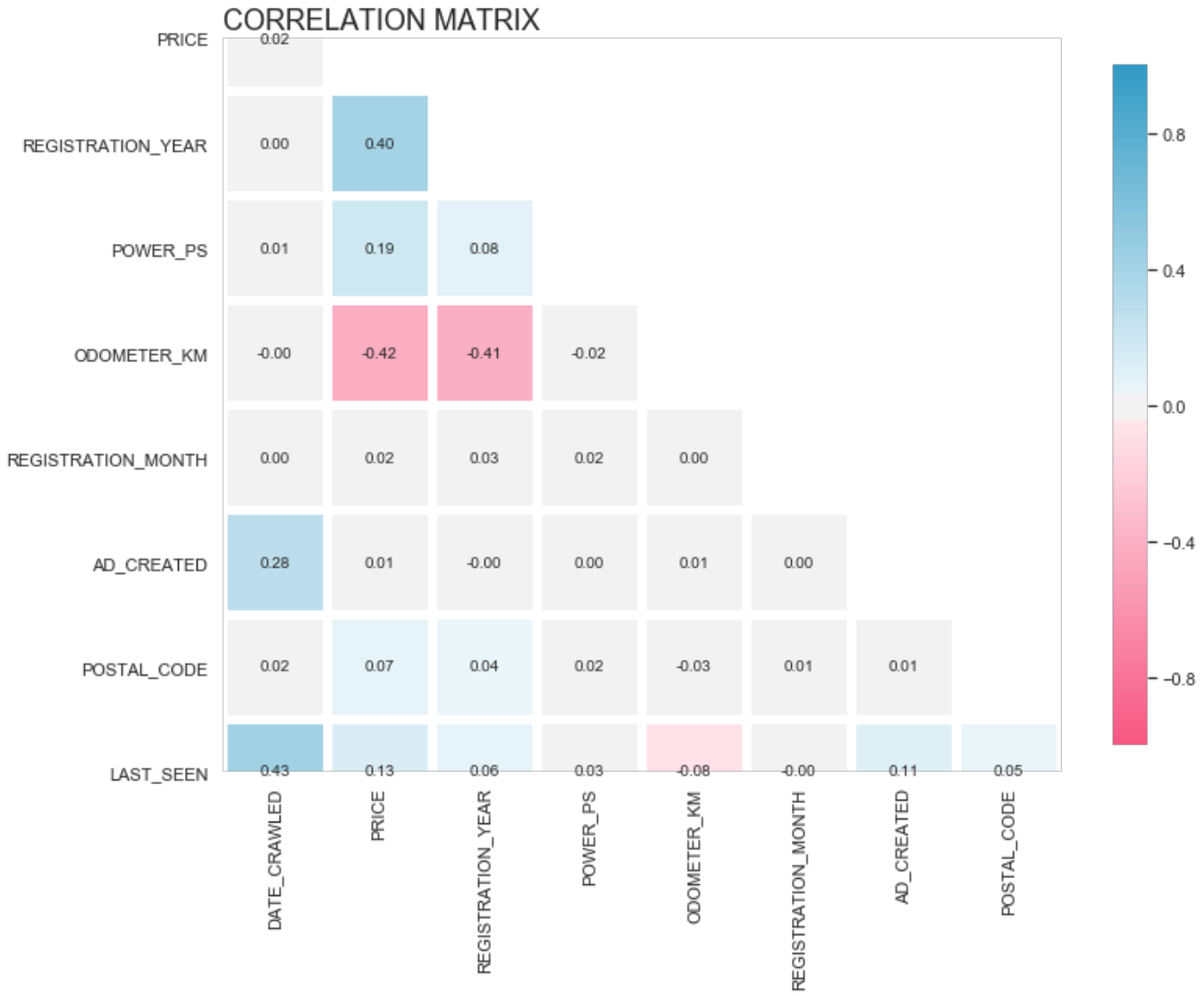
DATA SOURCE

Used Cars Database from Kaggle
https://www.kaggle.com/piumiu/used-cars-database-50000-data-points

Exploratory Data Analysis - Summary

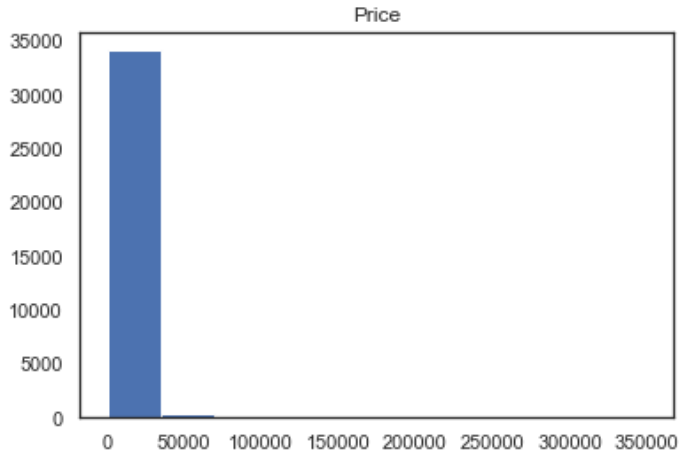
- The dataset consists of 20 columns
- 15 columns contain data of object type, 5 columns are int.64 type.
- 5 columns have missing values, but none of them contain more than 20% missing values
- Units of the variables are missing,

CORRELATION MATRIX



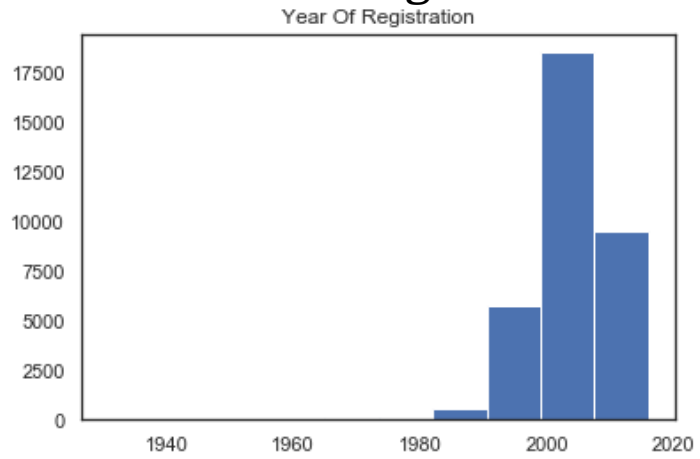
Distribution of Variables

Price



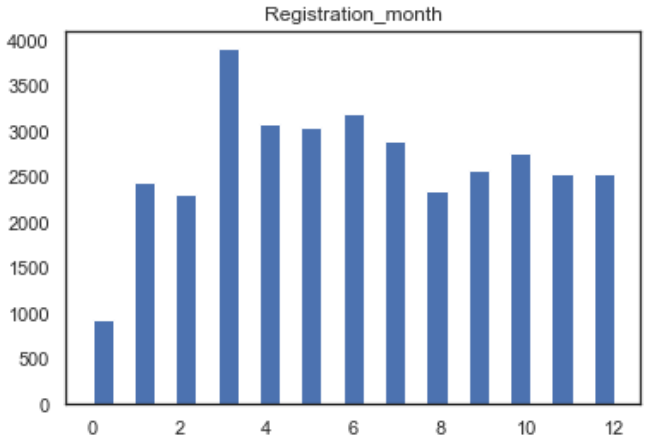
- Right skewed (positively skewed)

Years of Registration



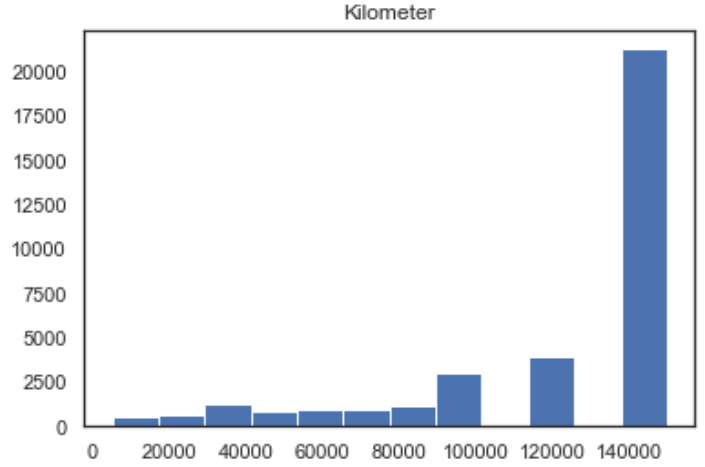
Left skewed (Negatively skewed)

Registration Month



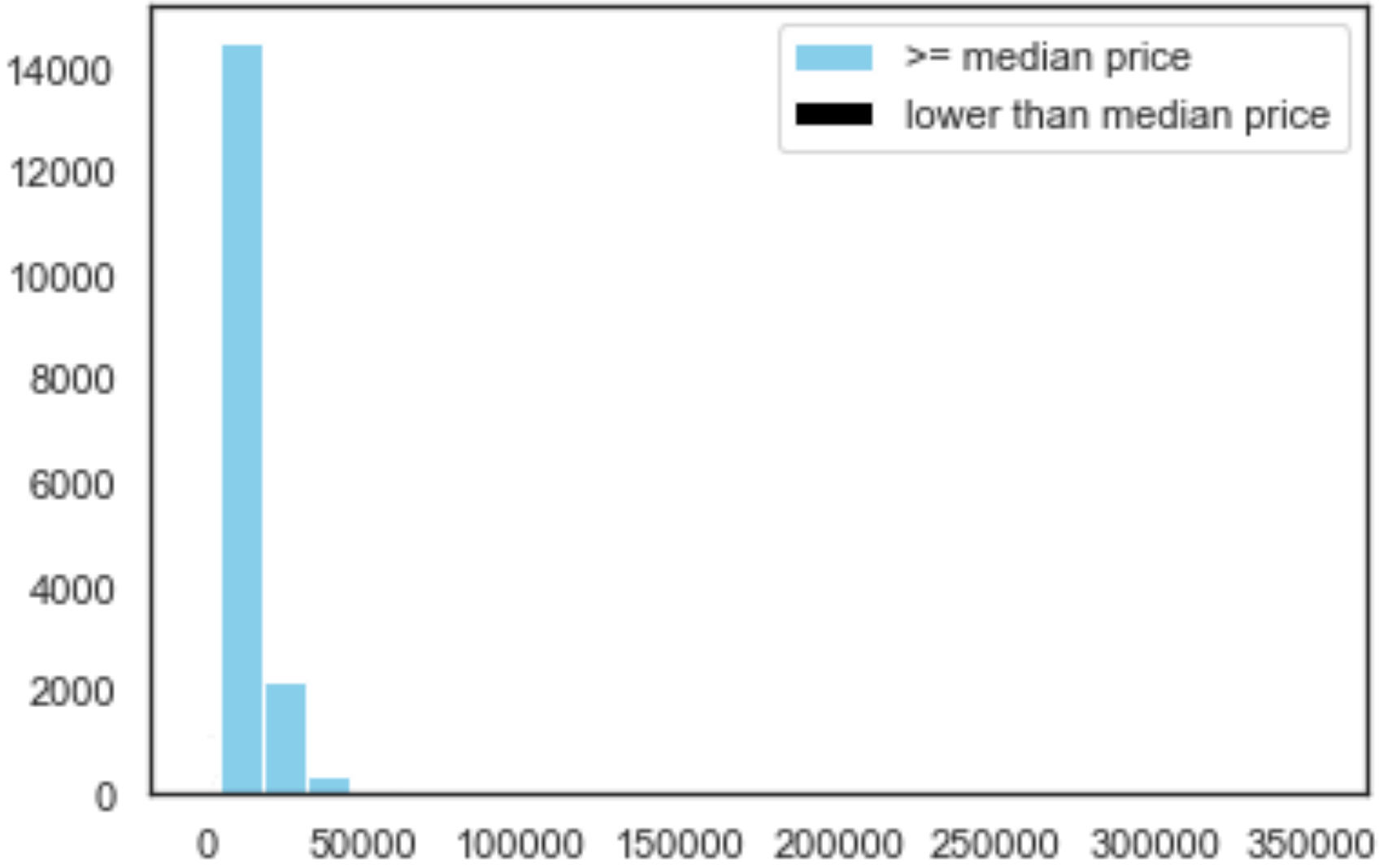
Normally Distributed

Odometer

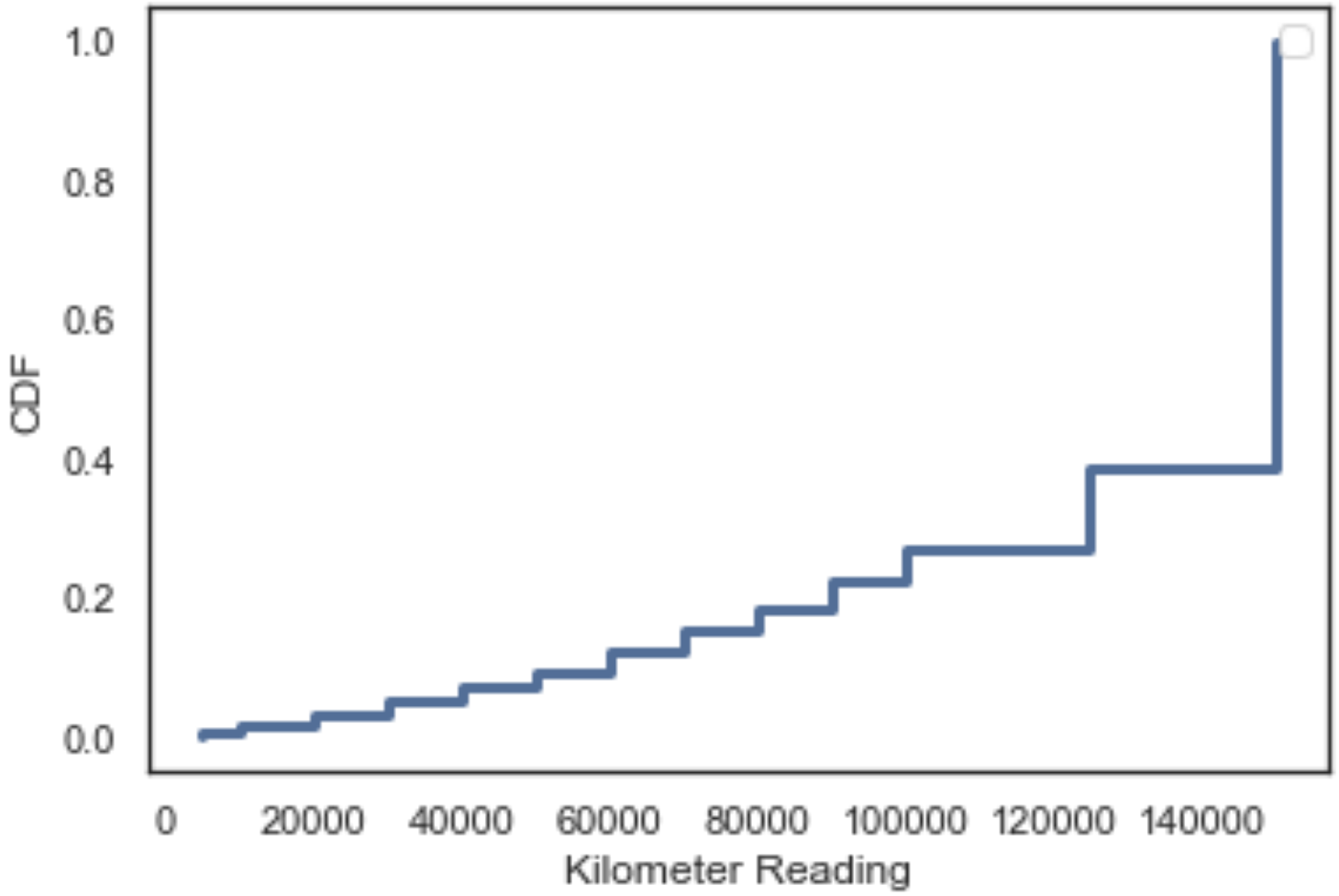


Left skewed (Negatively skewed)

Probability Mass Function (PMF)



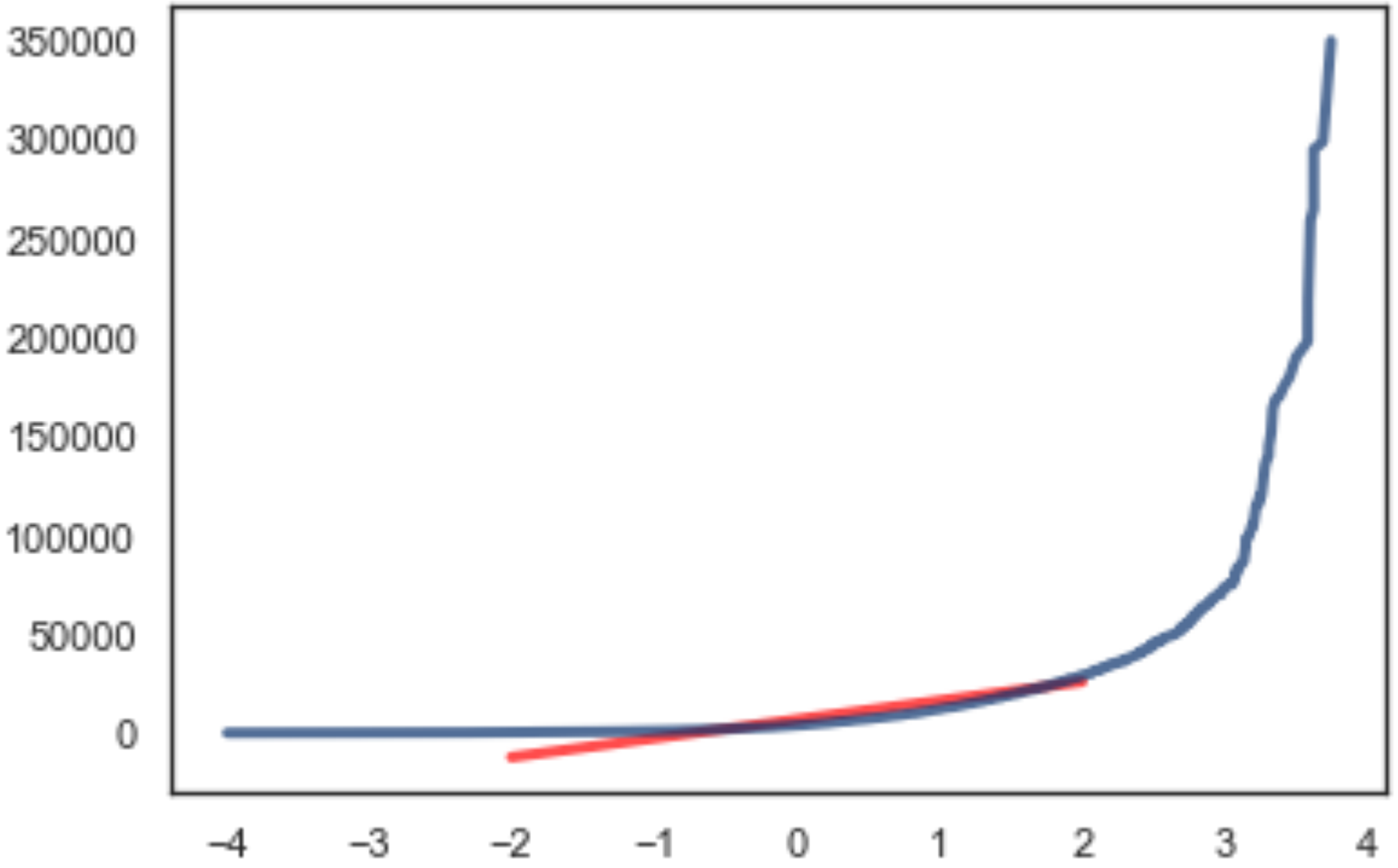
Cumulative Distribution Function (CDF)



- approx 25% cars have reading under 100000 km, and about 30% cars under 120000 kilometer.

- Common values appear vertical sections of the CDF; there are fewer values below 100000 kilometer, so the CDF in this range is flatter.

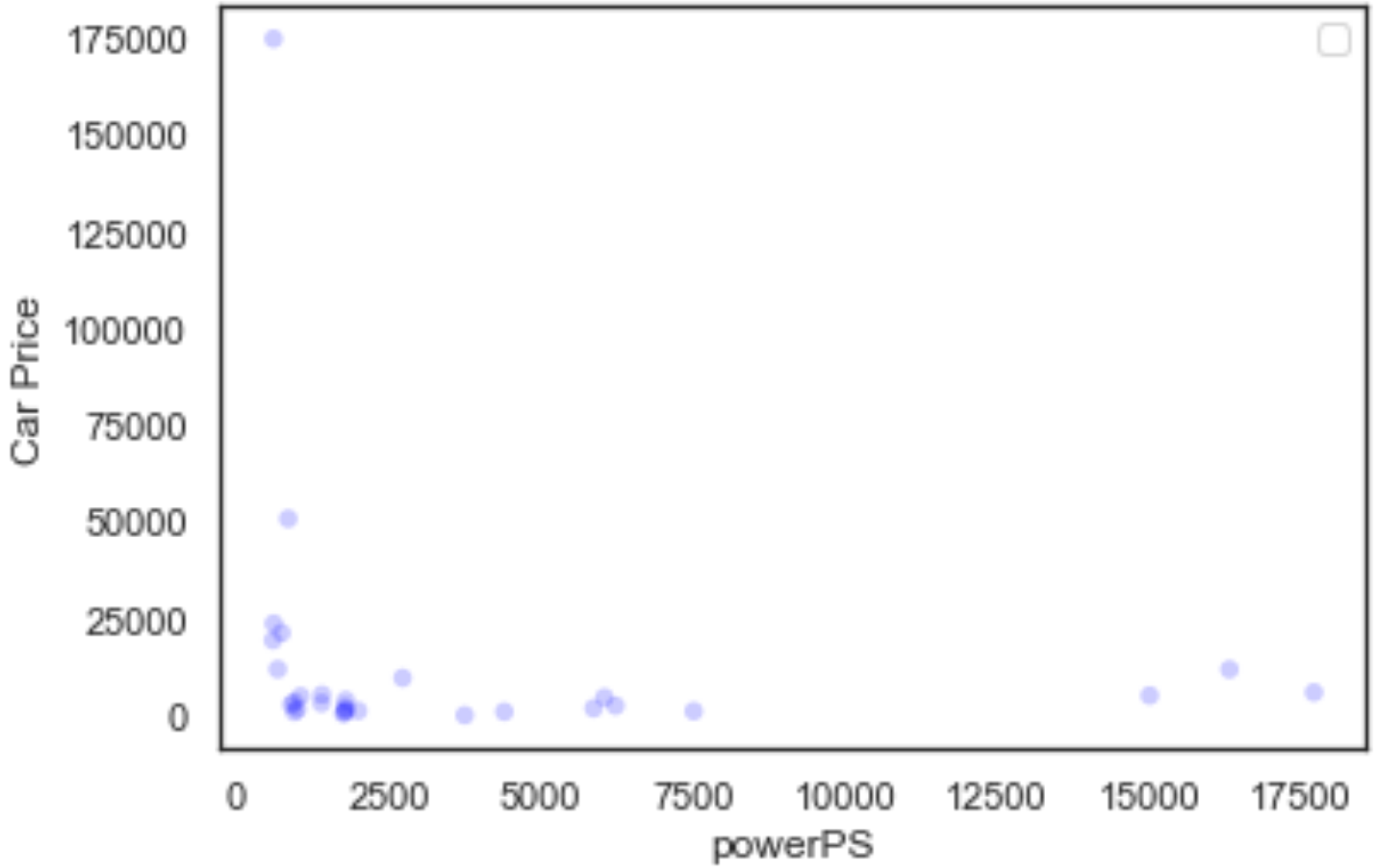
Analytical Distribution



Scatter Plot

Car price vs the power of the car in PS.

(Excluding low powerPS values)

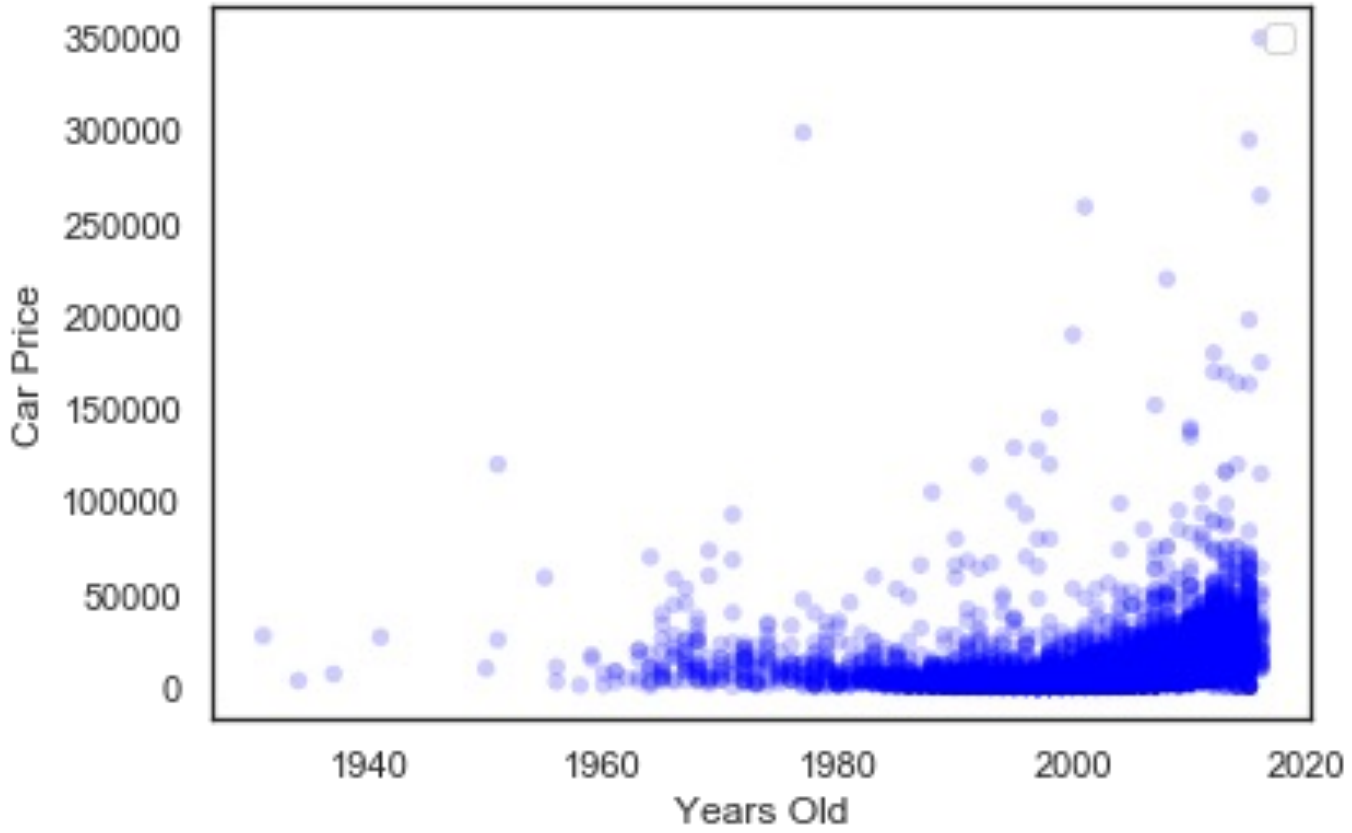


•powerPS and Car price have positive but weak correlation.

Scatter Plot

Age and price

(Excluded newer cars since their price is expected to be higher).



- Age is one of the factors that causes the price to change.
- Presence of few outliers for cars older than 35 years with high price range
- -Weak association of Car age and price.

Hypothesis Testing

Hypothesis Testing

Test correlation.

In [108]:
```
class CorrelationPermute(thinkstats2.HypothesisTest):

    def TestStatistic(self, data):
        xs, ys = data
        test_stat = abs(thinkstats2.Corr(xs, ys))
        return test_stat

    def RunModel(self):
        xs, ys = self.data
        xs = np.random.permutation(xs)
        return xs, ys
```

executed in 4ms, finished 16:09:45 2021-08-13

In [109]:
```
# Section data the two columns that we want to test
data = autos.price.values, autos.odometer_km.values

ht = CorrelationPermute(data)
ht.PValue()
```

executed in 877ms, finished 16:09:46 2021-08-13

Out[109]:
```
0.0
```

p-value `0` indicates a failure to reject the null hypothesis at the 5% significance level (p,0.05).

REGRESSION ANALYSIS

Regression Analysis

Regression Analysis of one dependent and multiple explanatory variables.

In [110]:
```
y = autos['price'] #value we are predicting - dependent variable
x = autos[['odometer_km', 'power_ps', 'registration_year']] #explanatory variables - Independent variables

X_train, X_test, y_train, y_test = train_test_split(x,y,test_size=0.30, random_state=10101)
#split the data 70/30

model = LinearRegression()
model.fit(X_train,y_train)
```

executed in 41ms, finished 16:09:49 2021-08-13

Out[110]:
```
LinearRegression(copy_X=True, fit_intercept=True, n_jobs=None, normalize=False)
```

In [111]:
```
coeff_df = pd.DataFrame(model.coef_, x.columns, columns=['coefficient'])
coeff_df
```

executed in 7ms, finished 16:09:51 2021-08-13

Out[111]:

	coefficient
odometer_km	-0.076433
power_ps	26.126443
registration_year	344.117862

As far as defination of *regression coefficient* concerned - it is the constant that represents the rate of change of dependent variable (price) as a function of changes in the independent variables (kilometer, powerPS, yearsOld, NoOfDaysOnline)

Thank you