Pilgrim Bank Case Study

Mason Hansen

1/22/2019

- Data Clean Up
  - Descriptive Statistics
  - t-tests for Profit
  - Linear Regression

Analysis of meaningfulness

- Business issue –Identification of core challenge
  - What is the business related assumption to be evaluated (Online usage impact on profitability)
- Statistical issue –Transformation of challenge to statistical test
  - How to transform business assumption to a statistical test (e.g. testable hypothesis)
- Statistical test
  - Selection -What specific statistical test is needed
  - Organization -What data extract to use, what software to use, how to setup data in software,
  - Execution -How to select/run the statistical analysis
- Statistical Result
  - Result interpretation –what is the statistical answer to the hypothesis test
- Business Conclusion
  - Business interpretation –What business conclusion can be drawn from the statistical answer

Case Formulation (Introduction) Q1. Based on the sample, what does customer profitability look like for the entire population?

- descriptive Statistics will be analyzed for customer profitability in order to gain better insight on potential variables that might have influence customer profitability for Pilgrim Bank in the year 1999. Q2. Challenge: Is online banking a beneficial channel that increases profitability or does it detract from profitability?
- Hypothesis:
- H_0: The average value of profitability for online banking and non-online banking customers is the same $(\mu\ Online\ = \mu\ NonOnline\ )$
- H_a: The average value of profitability for online banking and non_online banking customers is not equal $(\mu\ online\ \neq \mu\ non-online\ )$

Q3. What Role do Customer Demographics play in online versus offline customers?

- Statistical tests
  - A Student’s t-Test will be utilized to test the difference in mean profitability values between customers who use the online banking channel versus customers that do not use the online banking channel.

- A linear regression model will be applied to investigate customer demographics and their impact on customer profitability for the year 1999
- Odds ratios, calculated from the linear model will be investigated to highlight the influence of customer demographic variables on overall customer profitability

Data Clean Up

For the purpose of this case analysis, the data will come from Pilgrim Bank and include information about their customers from the years 1999 and 2000. This analysis will not use the data from 2000 and focus entirely on the 1999 data. The 1999 variables included in the analysis are as follows:

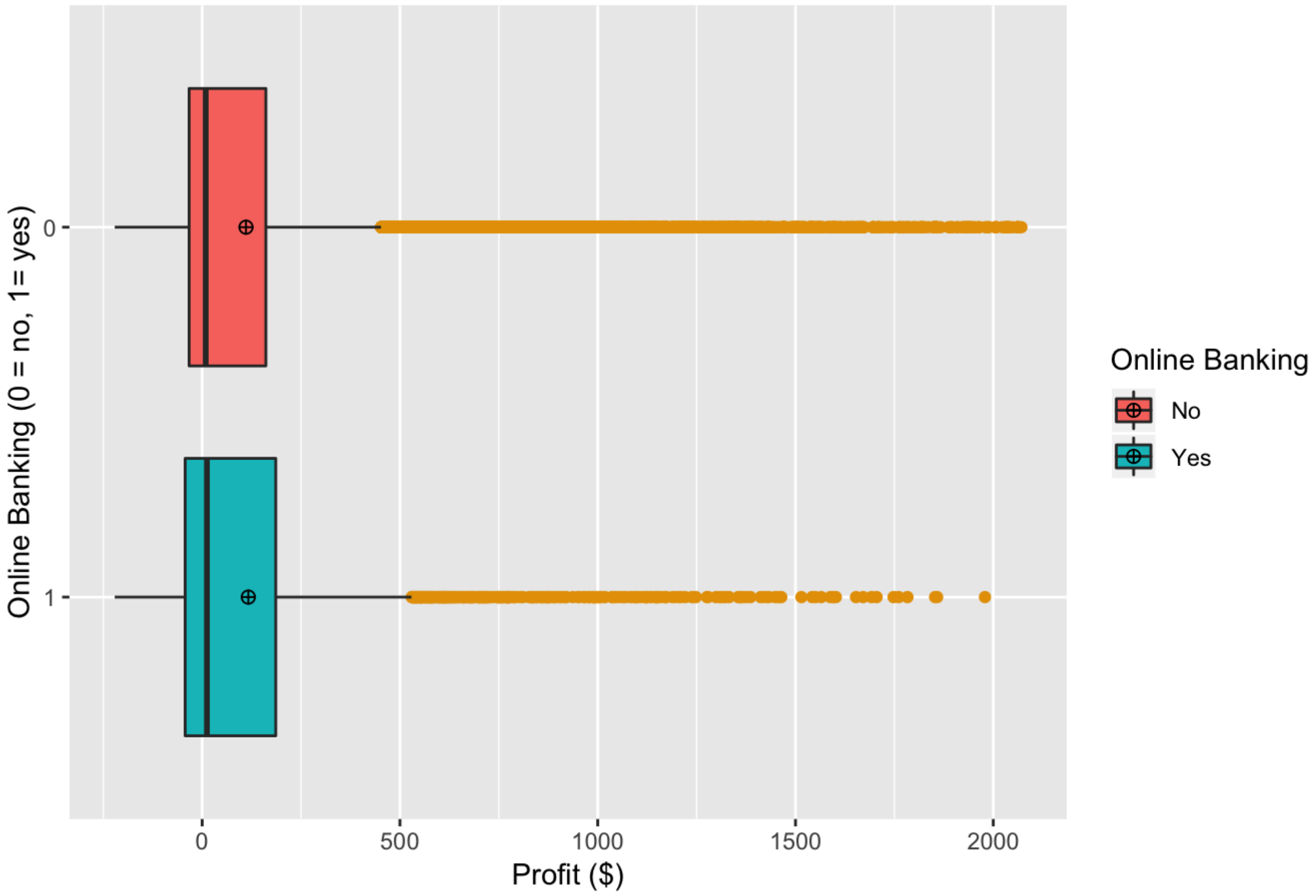
- Profit from customers (in US Dollars)
- Online Banking Usage (Yes / Online or No / Offline)
- Customer Age (Binned:)
- Customer Income(Binned:)
- Customer Loan Tenure (in years)
- Customer District
- Pay bills online (Yes / No)

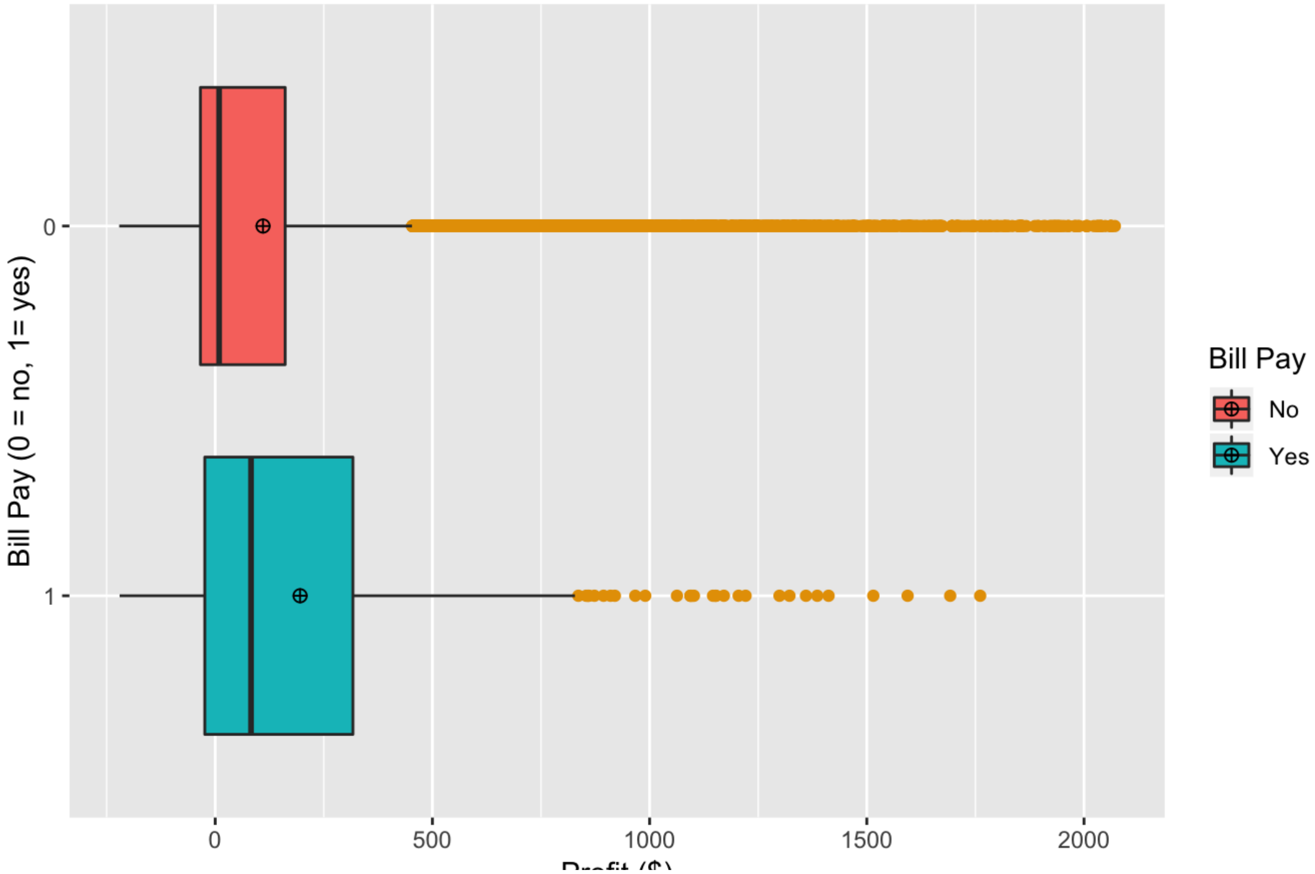
The data set is not entirely complete, including missing values in ~25% of the sample variables age and income. In order to account for missing values for proper analysis, deletion of ~25% of a sample is too aggressive. For this analysis, a dynamic method of imputation was chosen as the accurate replacement of missing values. The method use is called K-nearest_neighbors imputation (KNN). This method was chosen because all of the missing values were either age or income, which were previously binned into ranges of values. Since both age and income are now represented as categorical variables, the precision of a central tendency imputation of either the median or mean would seem inadequate. KNN is an algorithm that classifies each data point based on every independent variable in the data set (profit, age, income, tenure, district, online) and measures the distance from each data point to the central value (mean) of each cluster. For example, if person A and person B have similar profit, income and tenure, but only person A is missing their Age. The algorithm will calculate the distance between person A and person B, and filling in the missing value with the value of the “nearest neighbor”. In this scenario, person B is nearest to person A, so the missing age of person A will be imputed with person B’s age value. This algorithm can handle continuous, categorical and binary data, which is the main reason why this method was chosen.

Descriptive Statistics

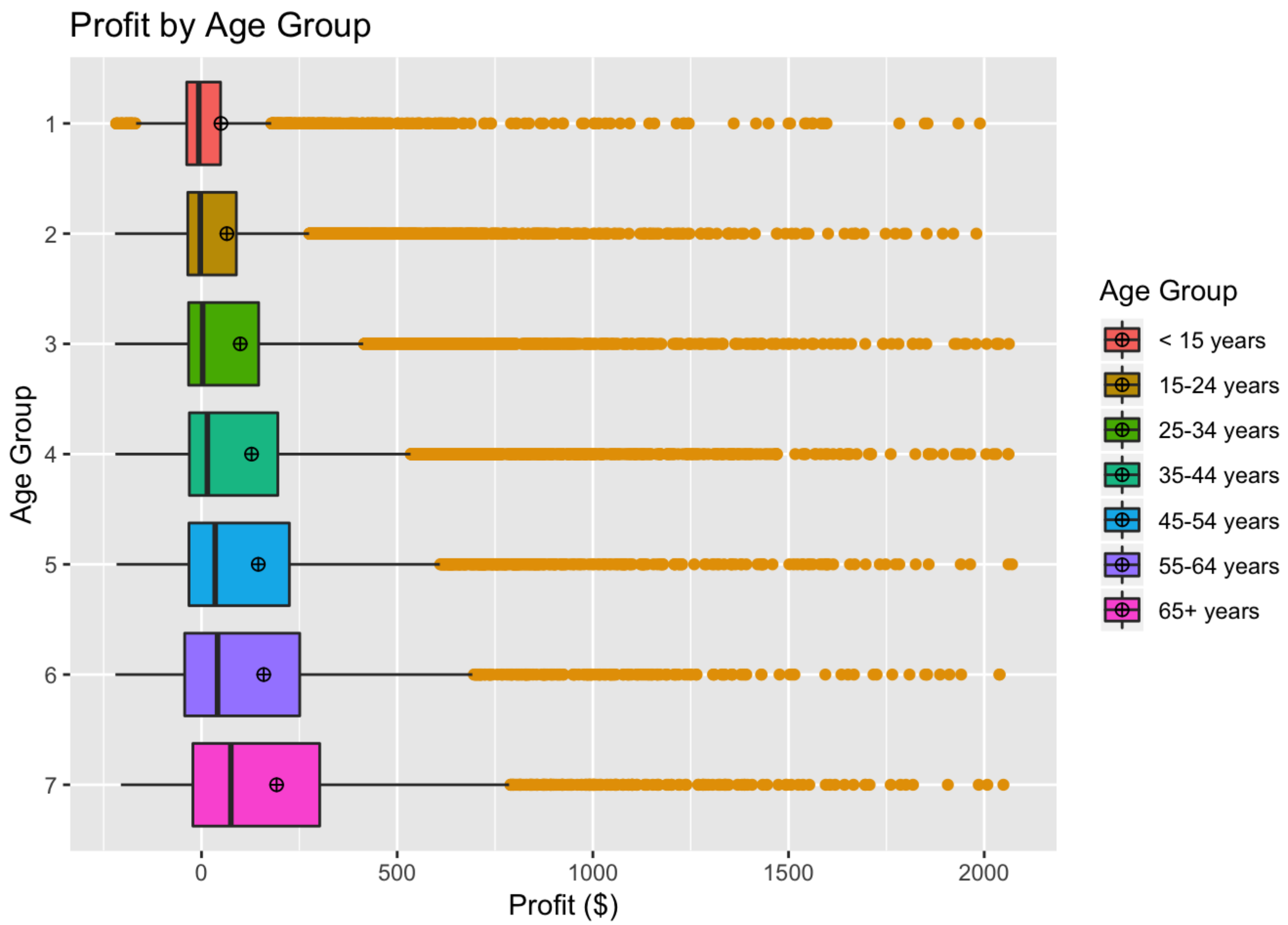
Descriptive statistics of the sample data set (shown below) shows that the average value of profit accross all customers is about 111.5. When accounting for customers that use online banking versus those who do not use online banking, the average value of profit for online customers is 116.67 and offline customers is 110.79. This might suggest that for customers that use the online banking channel generate sliglty more profit than those who do not use the online banking channel.

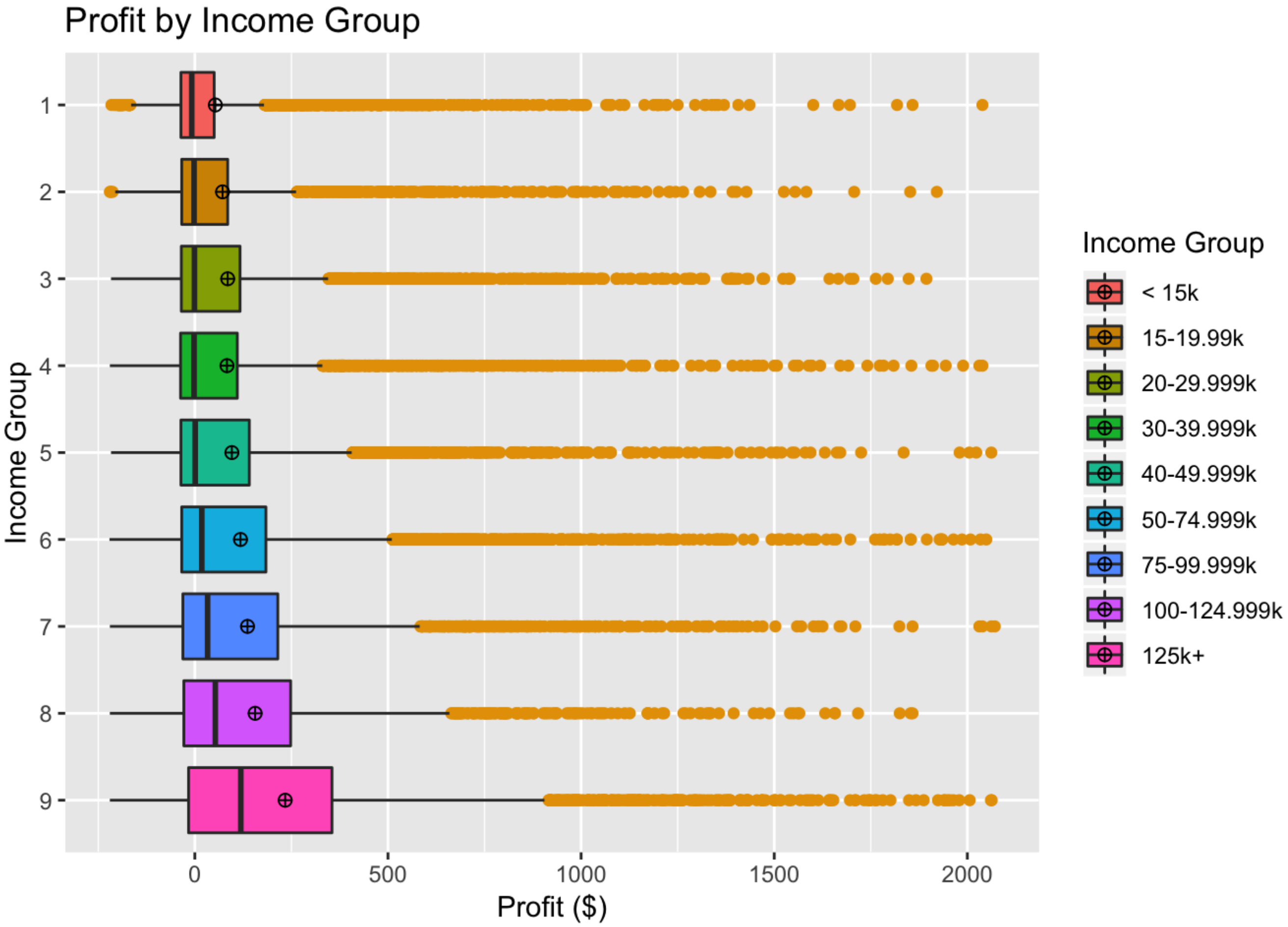
To better visualizle the data, a series of boxplots were plotted. Including each of variables Online, Age, Bill Pay and Income plotted against Profit. The variable District has been excluded from the visualization.

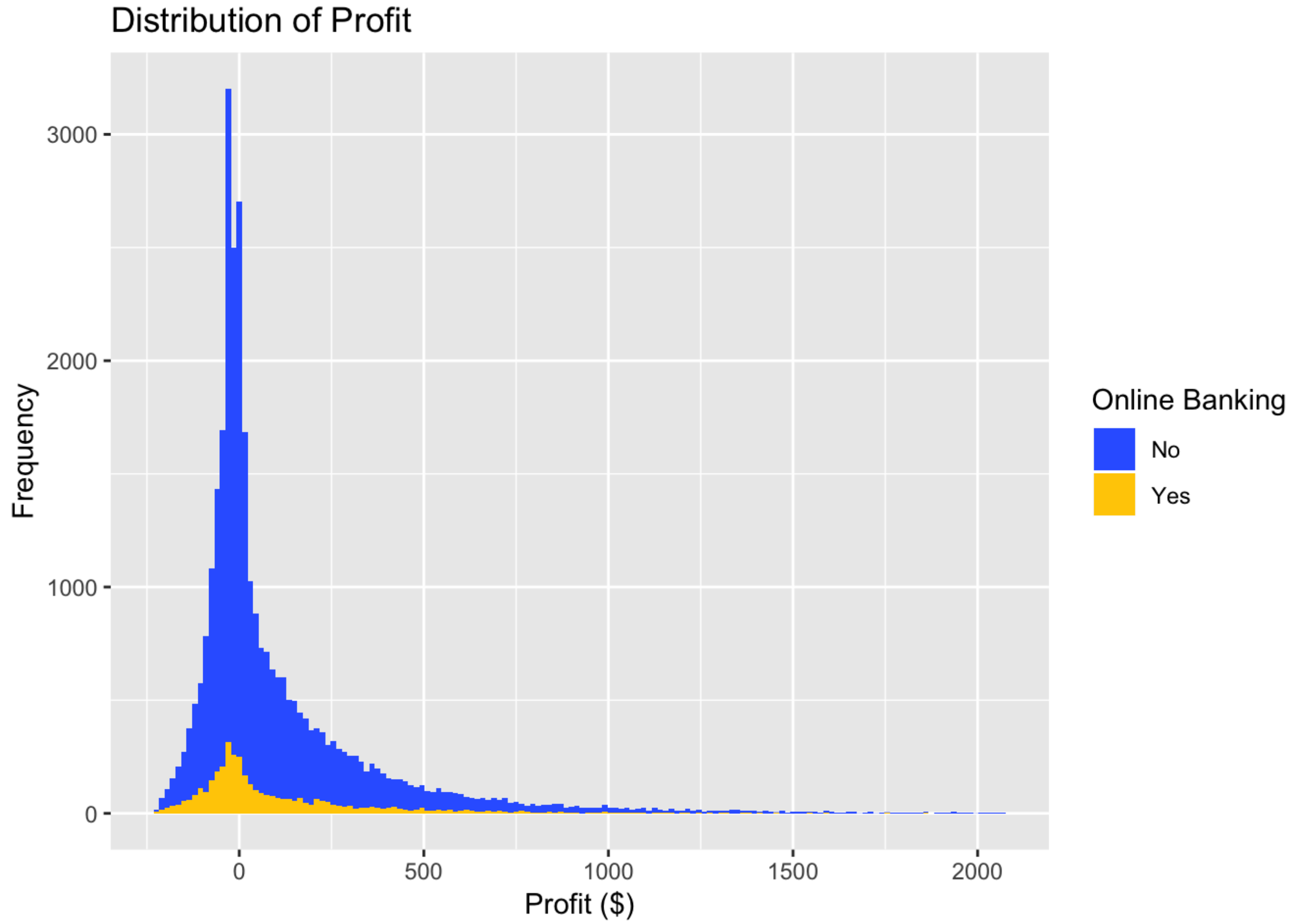

Profit of Online Banking


Profit of Online Bill Pay

Profit ($)


Profit by Age Group
Age Group
1
2
3
4
5
6
7
0
500
1000
1500
2000
Profit ($)
Age Group
< 15 years
15-24 years
25-34 years
35-44 years
45-54 years
55-64 years
65+ years


Profit by Income Group
Income Group
1
2
3
4
5
6
7
8
9
0
500
1000
1500
2000
Profit ($)
Income Group
< 15k
15-19.99k
20-29.999k
30-39.999k
40-49.999k
50-74.999k
75-99.999k
100-124.999k
125k+



t-tests for Profit

```
Welch Two Sample t-test
```

data: pilgrim_new$X9Profitbypilgrim_n ew$X9Online t = -1.2124, df = 4882.1, p-value = 0.2254 alternative hypothesis: true difference in means is not equal to 0 95 percent confidence interval: -15.389887 3.628706 sample estimates: mean in group 0 mean in group 1 110.7862 116.6668

Linear Regression

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-27.2	8.09	-3.36	0.000786
X9Billpay1	78.4	12.3	6.35	2.24e-10
X9District1200	17.7	5.12	3.45	0.00056
X9District1300	7.02	6.26	1.12	0.262
X9Inc2	18.4	7.46	2.46	0.0138
X9Inc3	26.3	6.18	4.26	2.09e-05

	Estimate	Std. Error	t value	Pr(>|t|)
X9Inc4	24.2	6.19	3.92	8.98e-05
X9Inc5	36.1	6.42	5.62	1.96e-08
X9Inc6	49.9	5.9	8.46	2.89e-17
X9Inc7	68.1	6.68	10.2	2.34e-24
X9Inc8	85	7.96	10.7	1.33e-26
X9Inc9	155	6.99	22.2	9.14e-109
X9Online1	0.863	4.89	0.176	0.86
X9Tenure	4.77	0.191	24.9	1.08e-135
X9Age2	4.16	6.3	0.66	0.509
X9Age3	19.8	6.26	3.16	0.00155
X9Age4	27.7	6.52	4.25	2.18e-05
X9Age5	35	7.19	4.87	1.13e-06
X9Age6	49.9	7.87	6.33	2.45e-10
X9Age7	84.5	7.68	11	3.89e-28