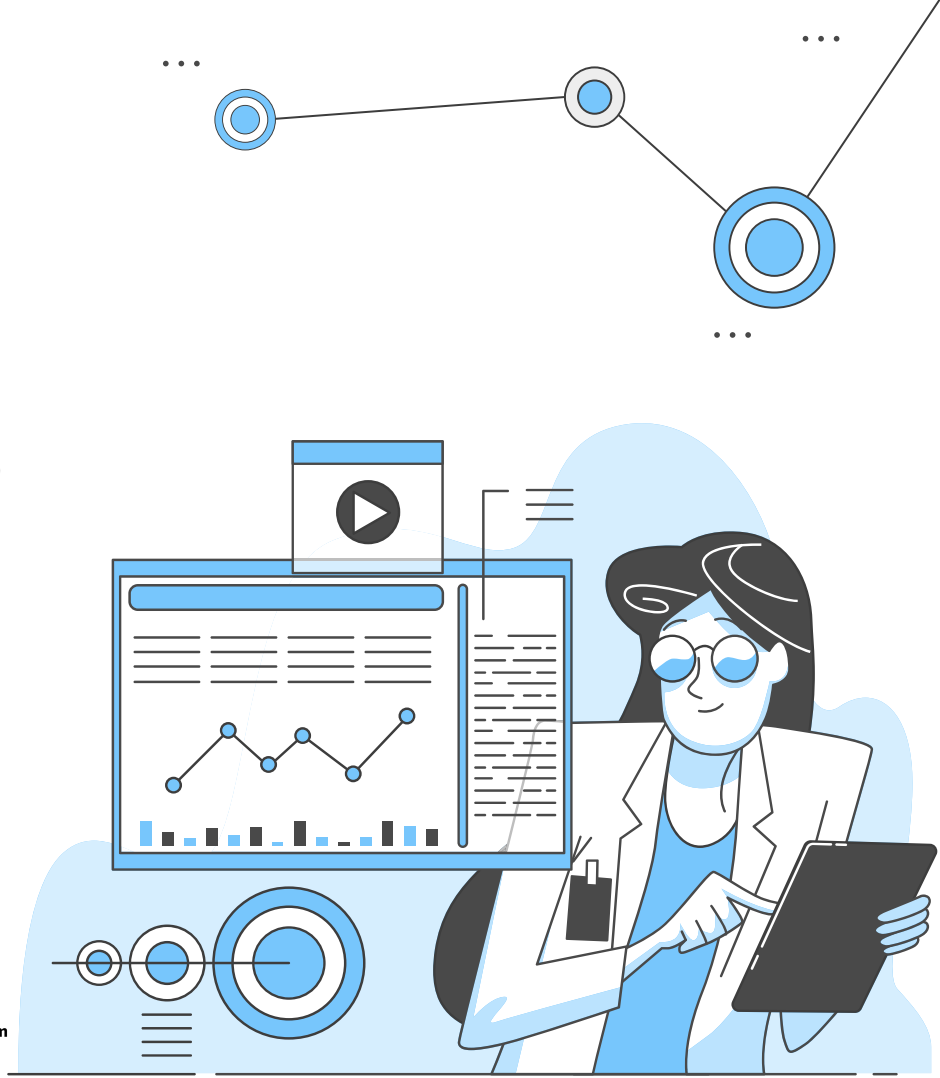
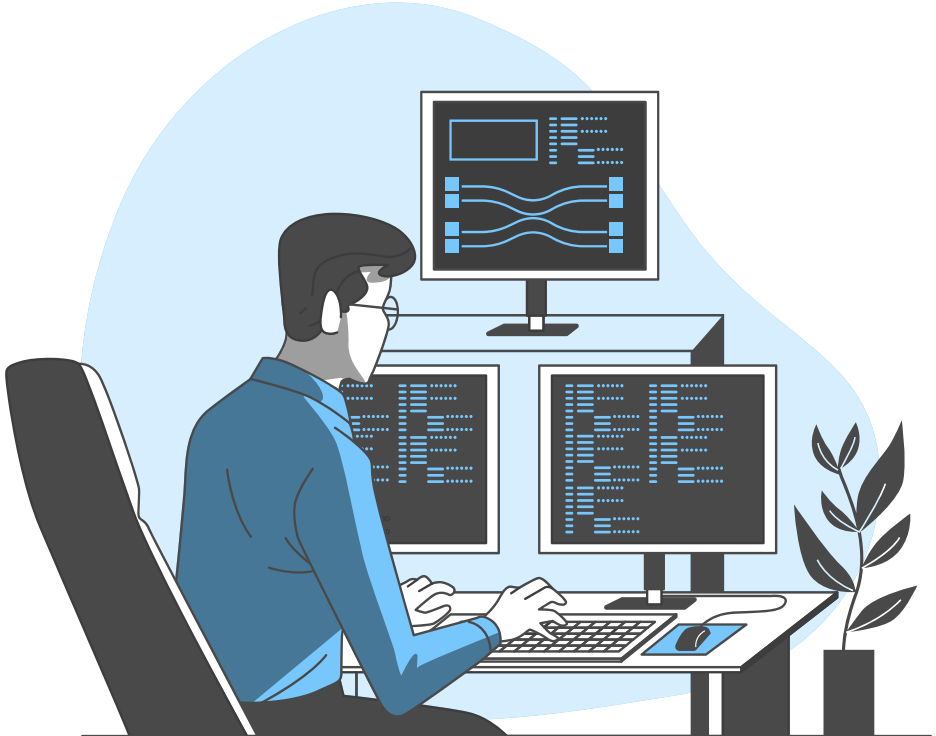
Bedding Bathing & Yonder : Sales Prediction



Francia F. Riesco
CSCI E-96 Spring 2022

HARVARD
Extension School
Professional Development Program

Today Agenda



01

Case Introduction

Predictive model data analysis

02

Exploratory Data Analysis

Review training, test and prospective dataset

03

Predictive Modeling

Linear Regression, KNN, RandomForest

04

Results

Which model is both accurate and consistent

01

Introduction

Predictive model data analysis

BBY

Bedding Bathing & Yonder

BBY is an American chain of domestic merchandise retail stores with an online presence. The chain primarily operates stores throughout the United States.

02

Exploratory Data Analysis

Fields

Categorical

- Donate Environment Cause In Home
- Donate To Charity In Home
- Residence HH Gender Description
- EthnicDescription
- BroadEthnicGroupings
- Presence Of Children Code
- HomeOwner Renter
- Media Education Years
- Education
- Occupation Industry
- ComputerOwnerInHome
- FirstName
- Last Name
- Gender
- Telephone FreePhone
- county
- city
- state
- Dwelling Unit Size
- store Visit Frequency
- PropertyType
- Parties Description
- Religions Description
- Gun Owner
- Veteran

Numeric

- tmpID
- lat
- lon
- Age
- NetWorth
- fips
- state Fips
- Land Value
- EstHomeValue
- ISPSA

Dependent Variable

- yHat The average household spend with BBY in USD

fields

Discarded incomplete Fields

- Religious Contribution In Home
- PoliticalContributerInHome
- Donate to Animal Welfare
- Donate to Arts and Culture
- Donates Children Causes
- Donate to Healthcare
- DonatestoInternationalAidCauses
- Donate to Veterans Causes
- Donate to Healthcare 1
- DonatestoInternationalAidCauses1
- Donate to Wildlife Preservation
- DonatestoLocalCommunity
- Mosaic Z4
- Investor
- Business Owner
- Horse Owner
- CatOwner
- Dog Owner
- OtherPetOwner
- HomeOffice
- BookBuyerInHome
- Upscale Buyer Home
- Buyer of Antiques in Household
- BuyerofArtinHousehold
- GeneralCollectorinHousehold
- BooksAudioReadinginHousehold
- Home Purchase Price
- Family Magazine Home
- Female Oriented Magazines In Home
- Religious Magazine In Home
- Gardening MagazineS Home
- Culinary Interest Magazine In Home
- Health Fitness Magazine In Home
- Do It Yourself Magazine Home
- Financial Magazine In Home
- Interest in Current Affairs Politics In Household
- Likely Union Member
- supports Affordable Care Act
- supportsGayMarriage
- supports Gun Control
- supports Taxes Raise
- overall social views
- DonatestoConservativeCauses
- DonatestoLiberalCcauses

Dataset by Numbers

15K

Training Set

5K

Test Set

~6K

Prospects Set

80

Fields

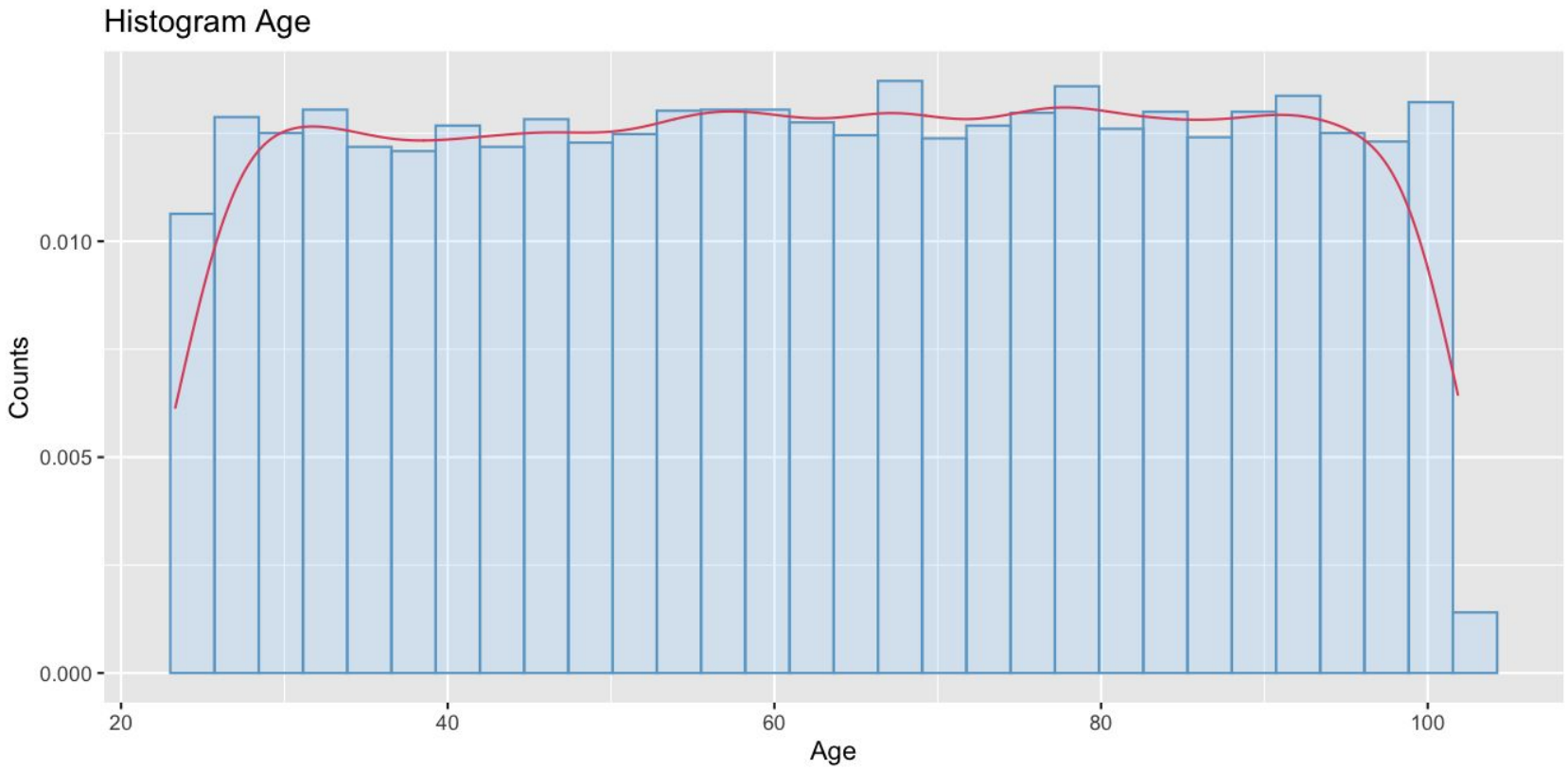
31

Independent variables

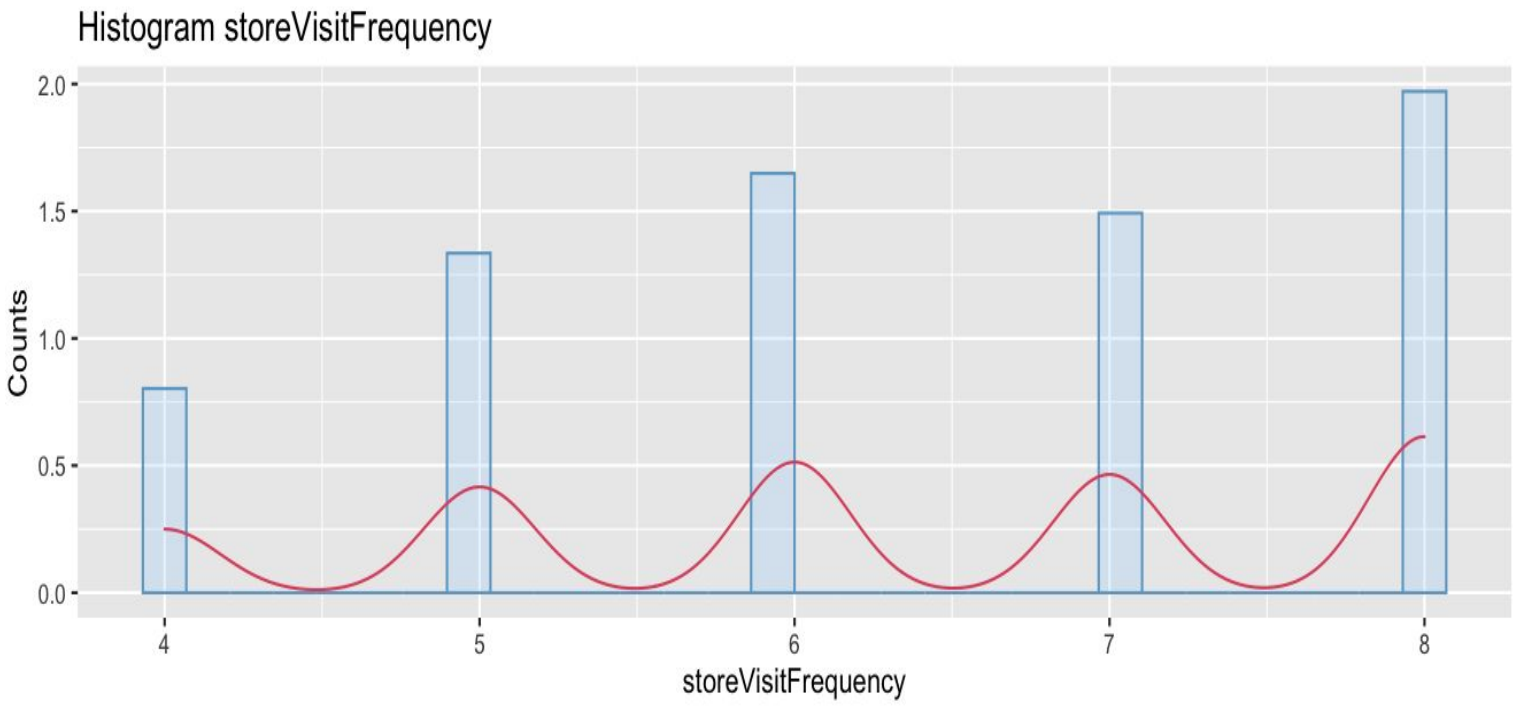
1

Dependant variable

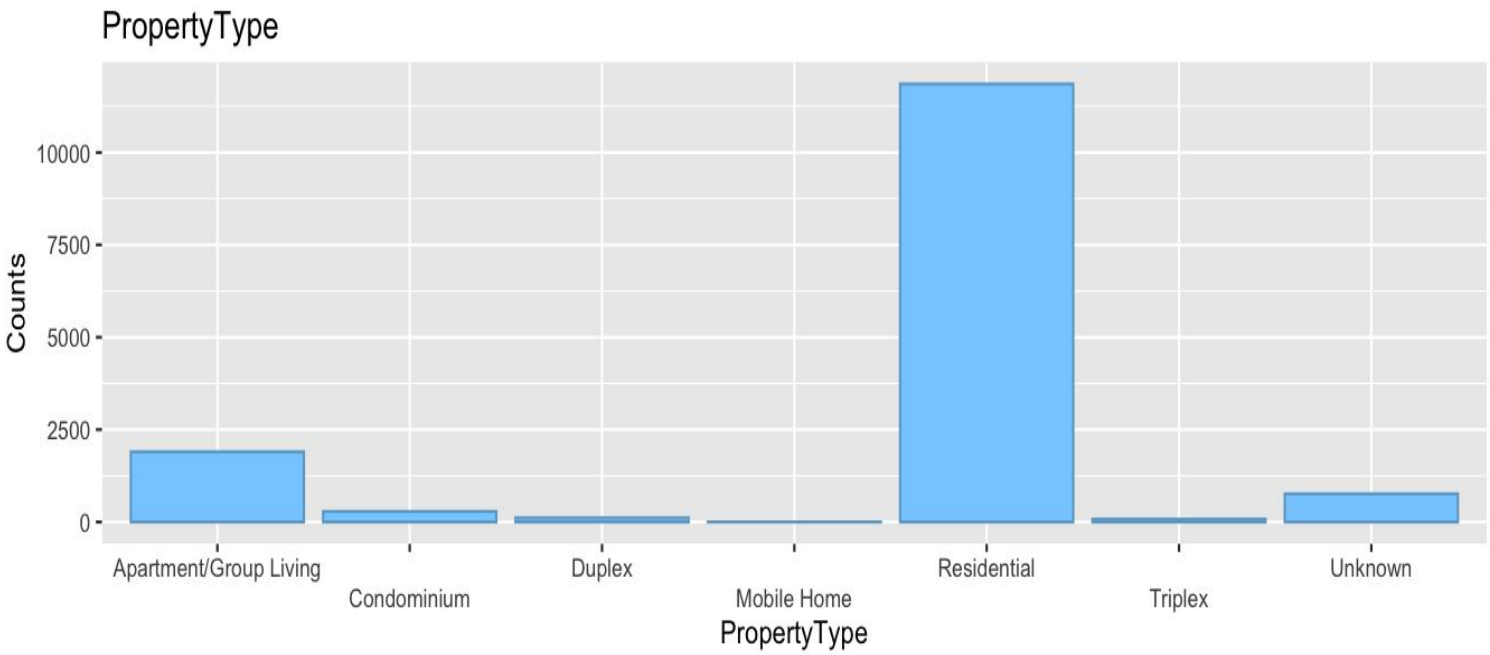
Exploring dataset



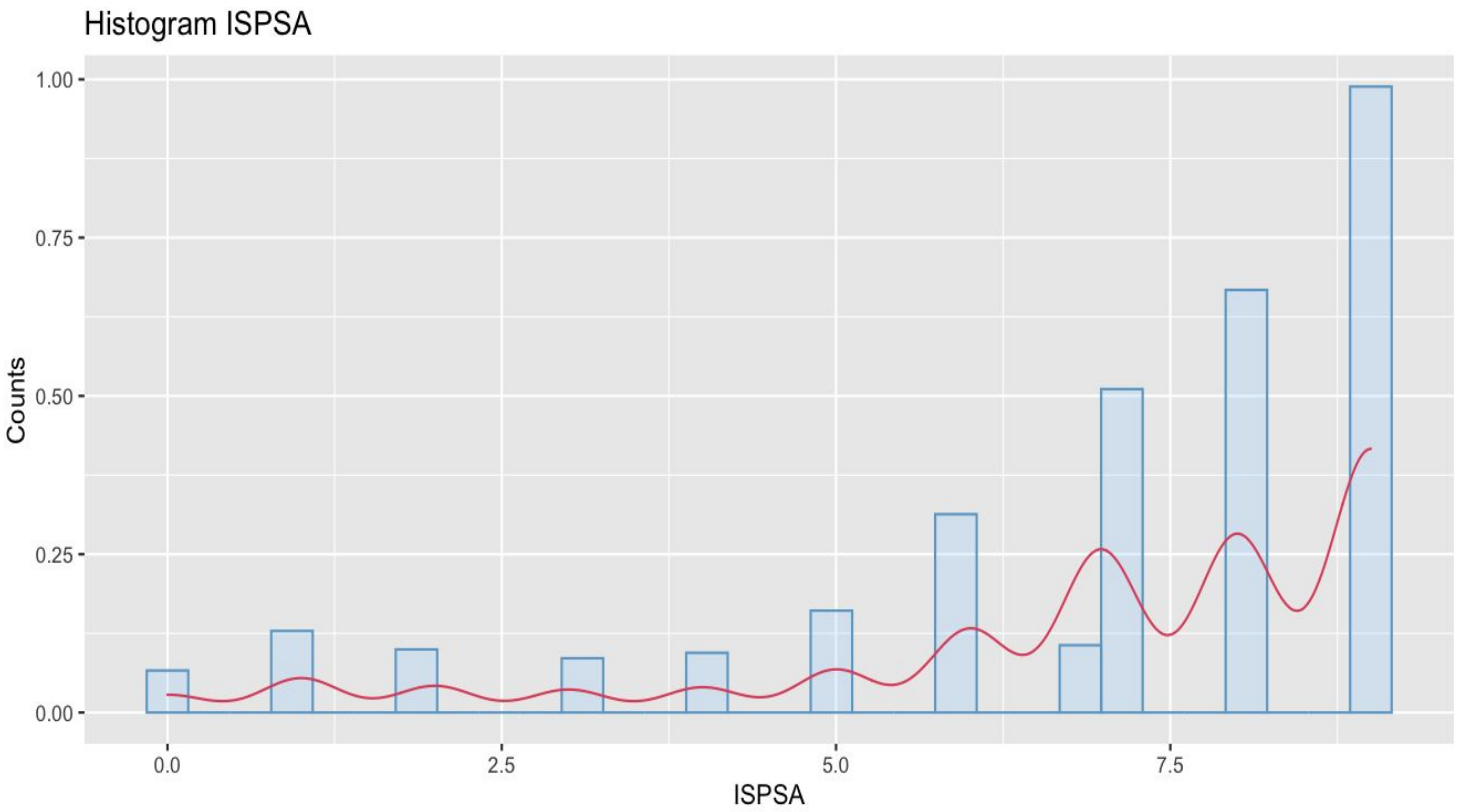
Exploring dataset



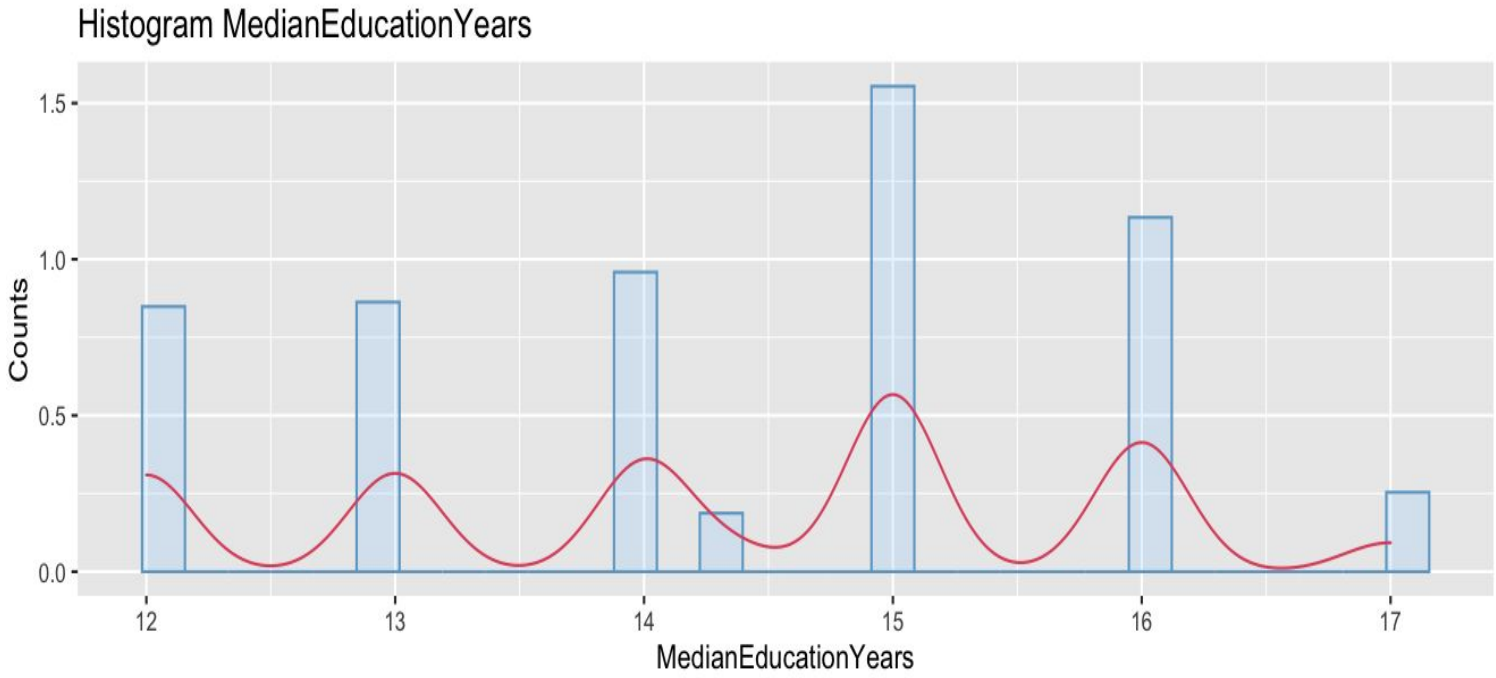
Exploring dataset



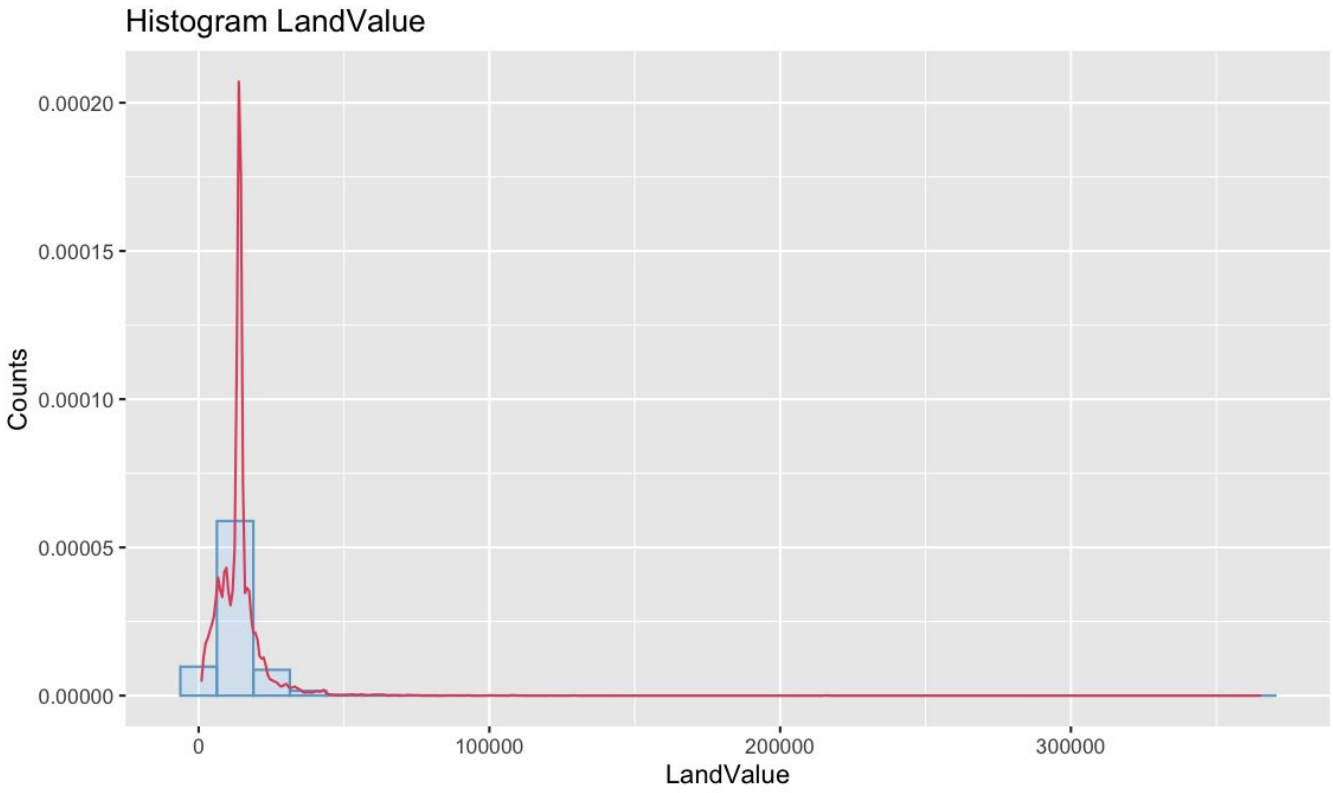
Exploring dataset



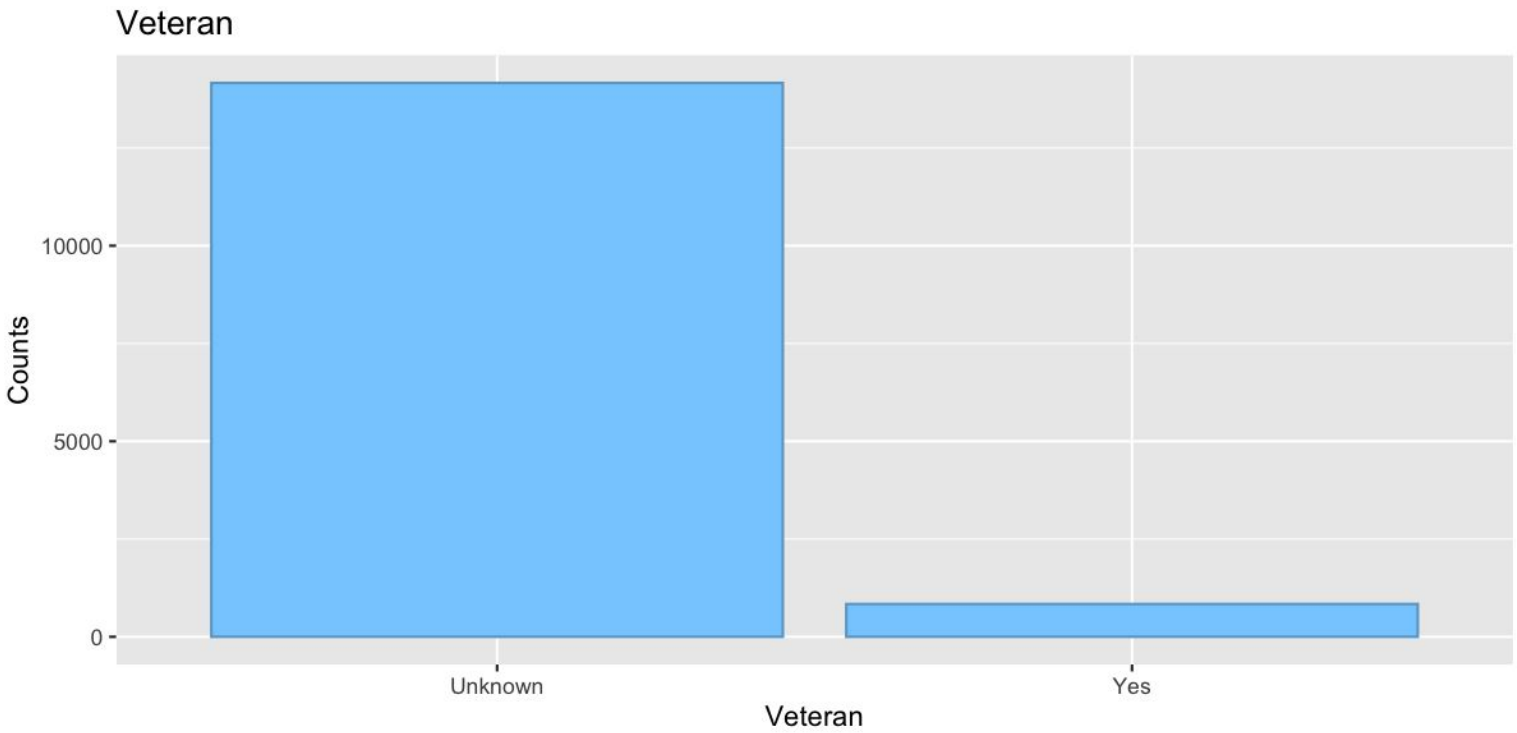
Exploring dataset



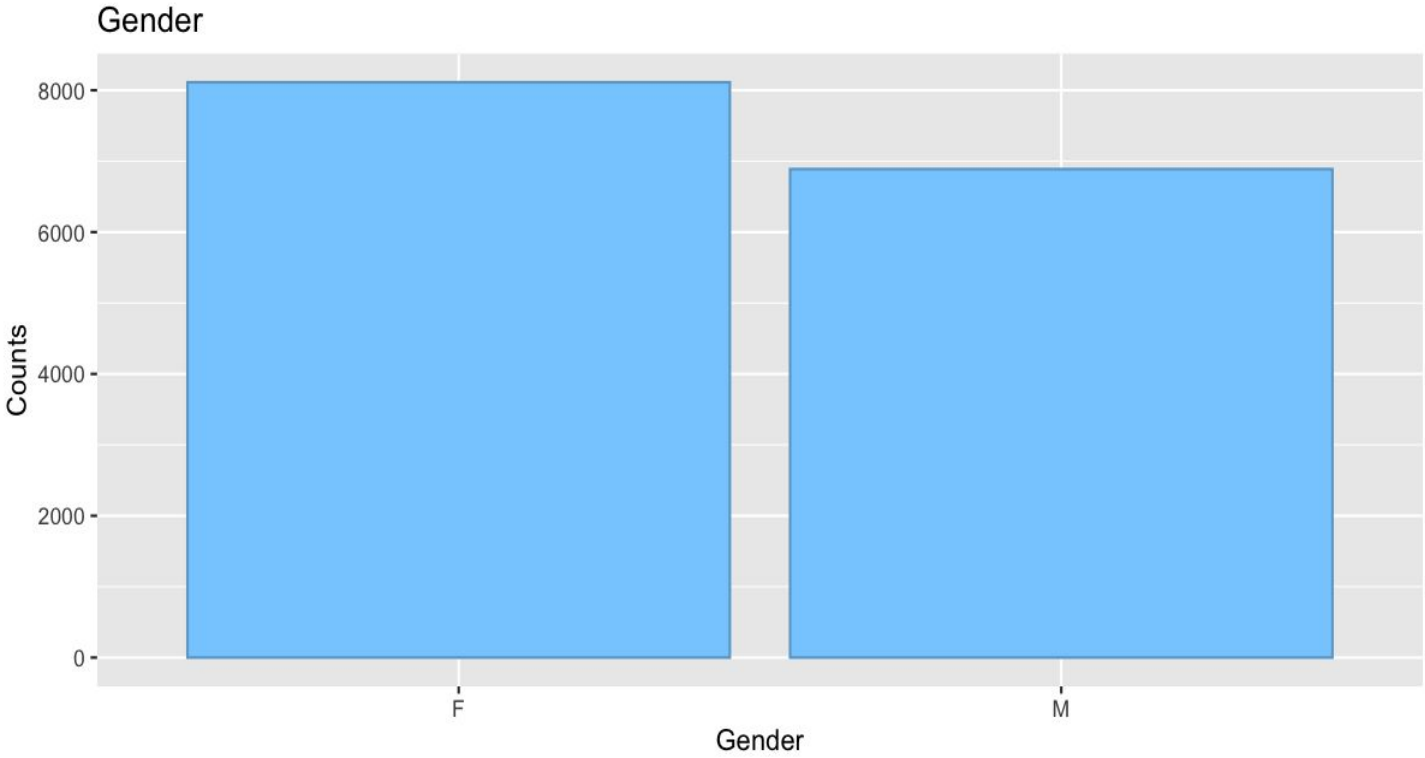
Exploring dataset



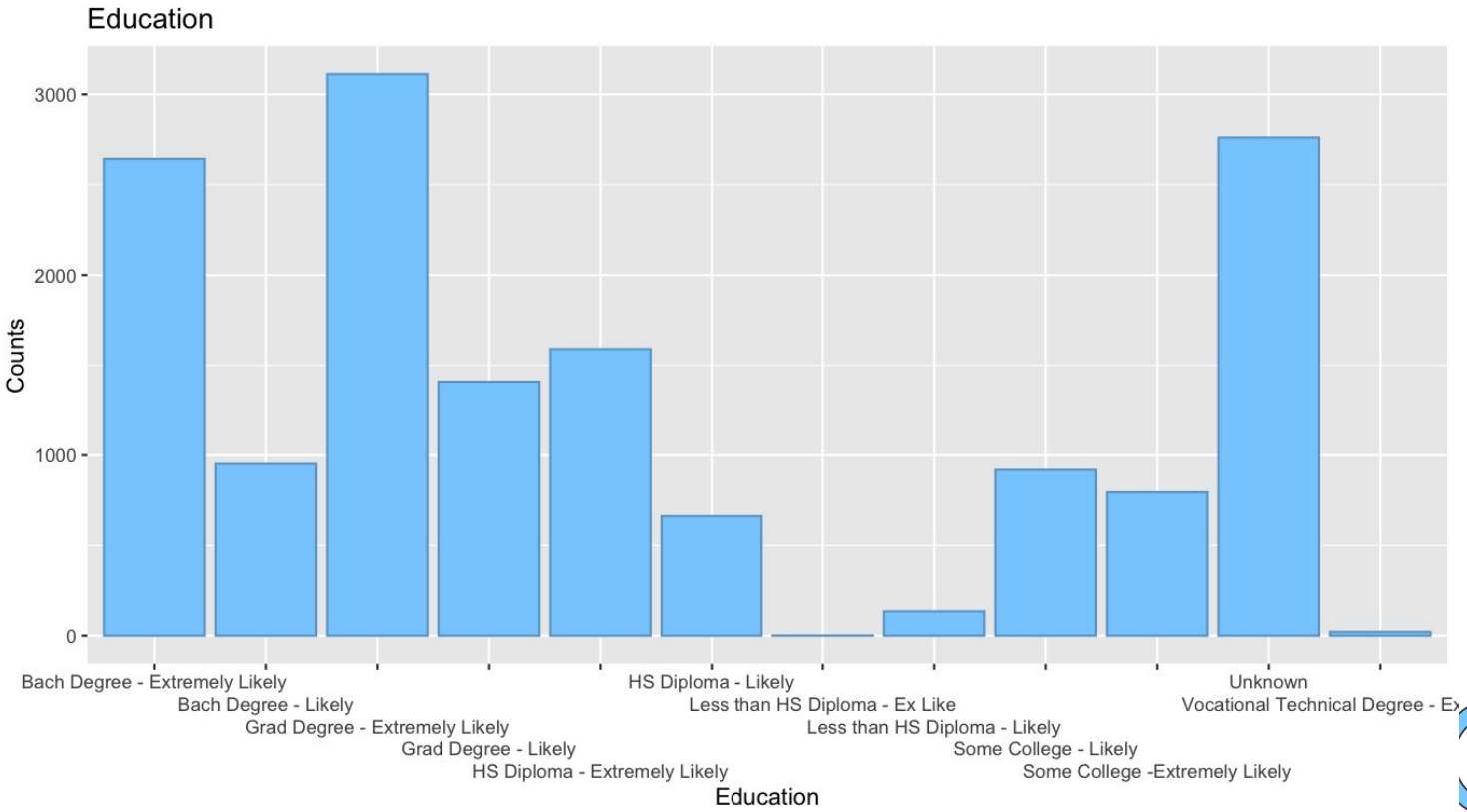
Exploring dataset



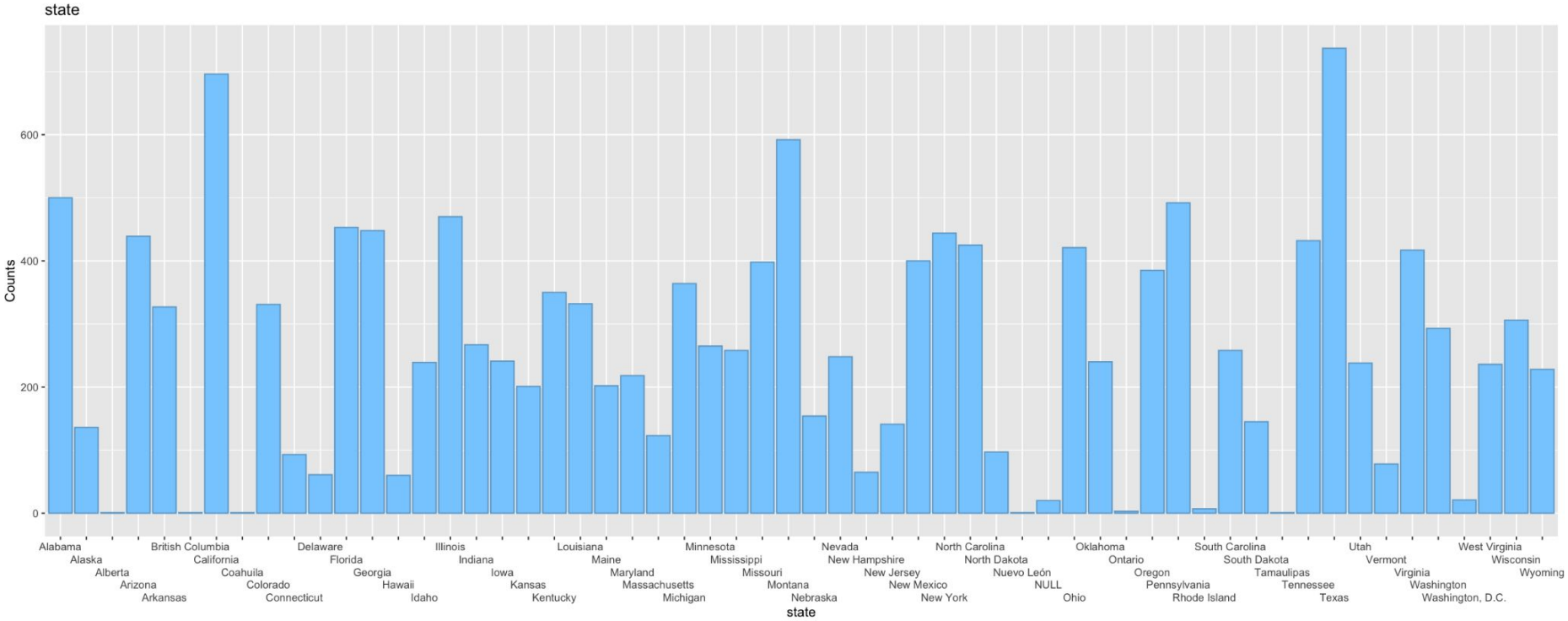
Exploring dataset



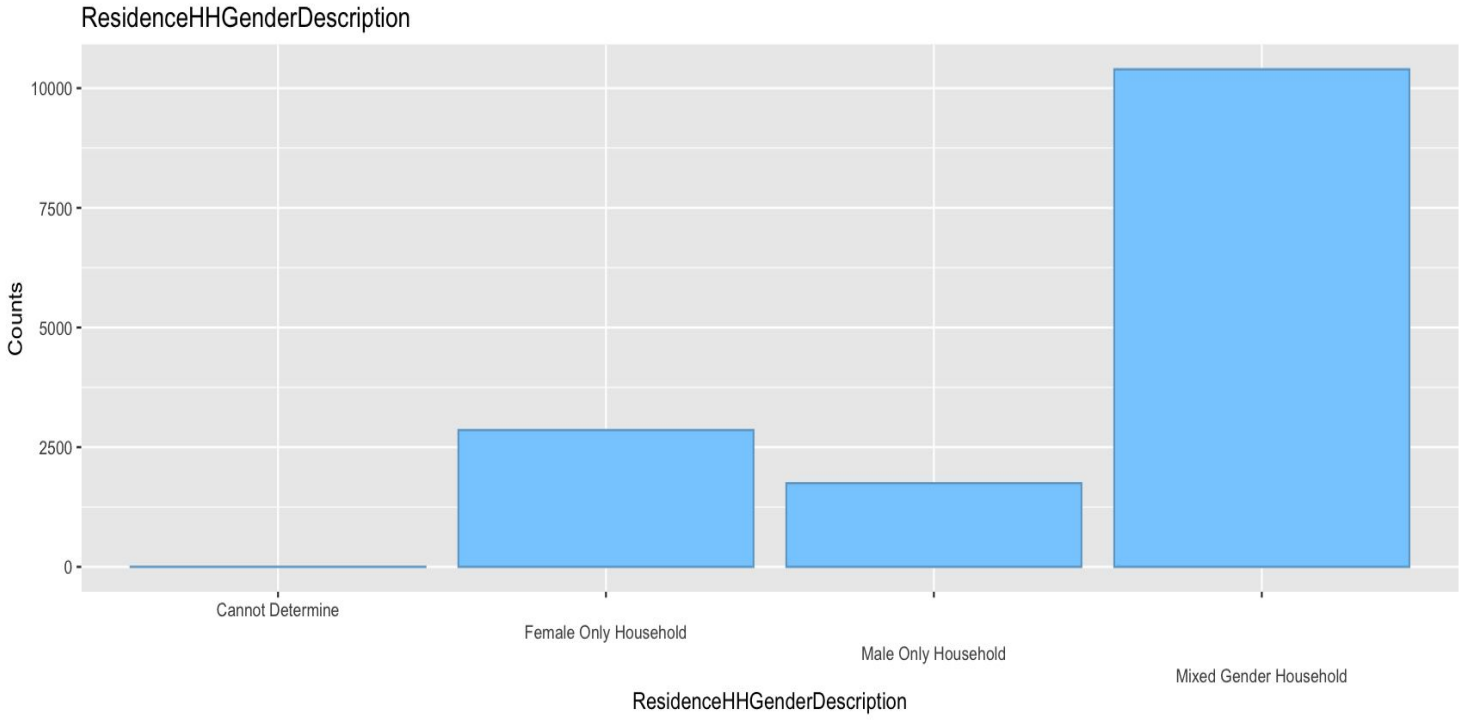
Exploring dataset



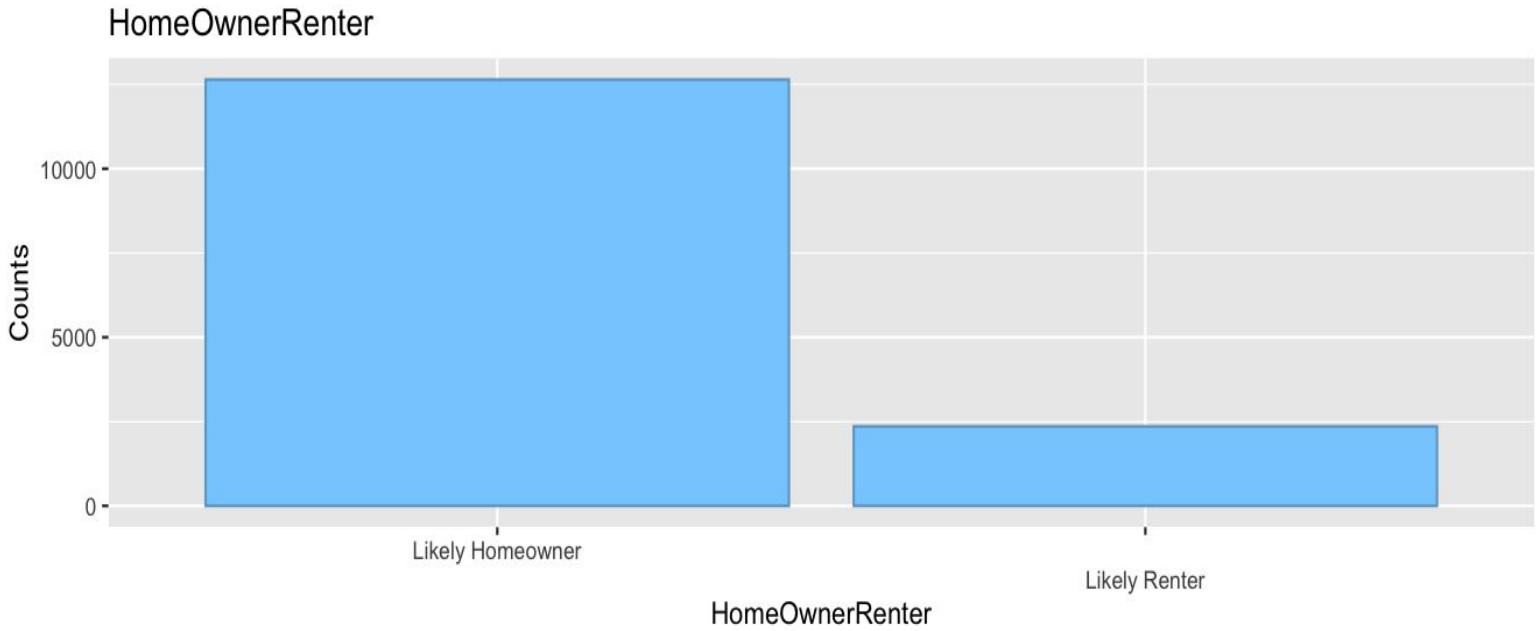
Exploring dataset



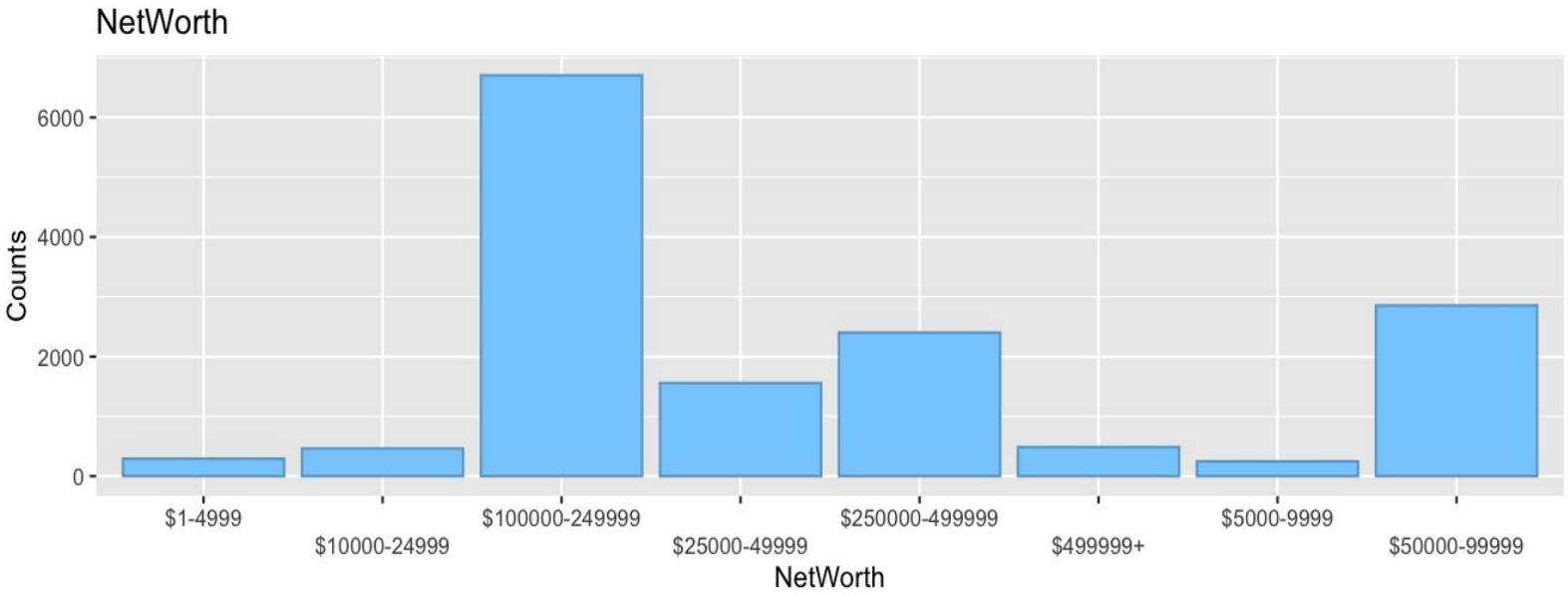
Exploring dataset



Exploring dataset



Exploring dataset



03

Predictive Modeling

Modeling Goals

Residuals

The shape of the residuals histogram is almost symmetric. If it is not symmetric our hypothesis assumption has been violated, and our model fails.

Hypotheses

All the Informative variables are significant predictor of the spend yHat variable

Key Performance Indicator

ME: Mean Error
RMSE: Root Mean Squared Error
MAE: Mean Absolute Error
MPE: Mean Percentage Error
MAPE: Mean Absolute Percentage Error

Residuals vs Fitters

When model is suitable for a data set, then the residuals are more or less randomly distributed around the 0 line

Main Libraries

Caret
MLmetrics
ggplot

KPI Performance

Low RMSE, high R^2 *
Low RMSE, low R^2
High RMSE, high R^2
High RMSE, low R^2 **

* best case ** worst case

The Models

Linear Regression

A linear regression analyzes the relationship between a response variable and one or more variables.

Random Forest

It builds the multiple decision trees which are known as forest and glue them together to urge a more accurate and stable prediction

k-Nearest Neighbors

KNN is a method for estimating the likelihood that a data point will become a member of one group based on the nearest point that it belong

Linear Regression

72.98131
Train RMSE

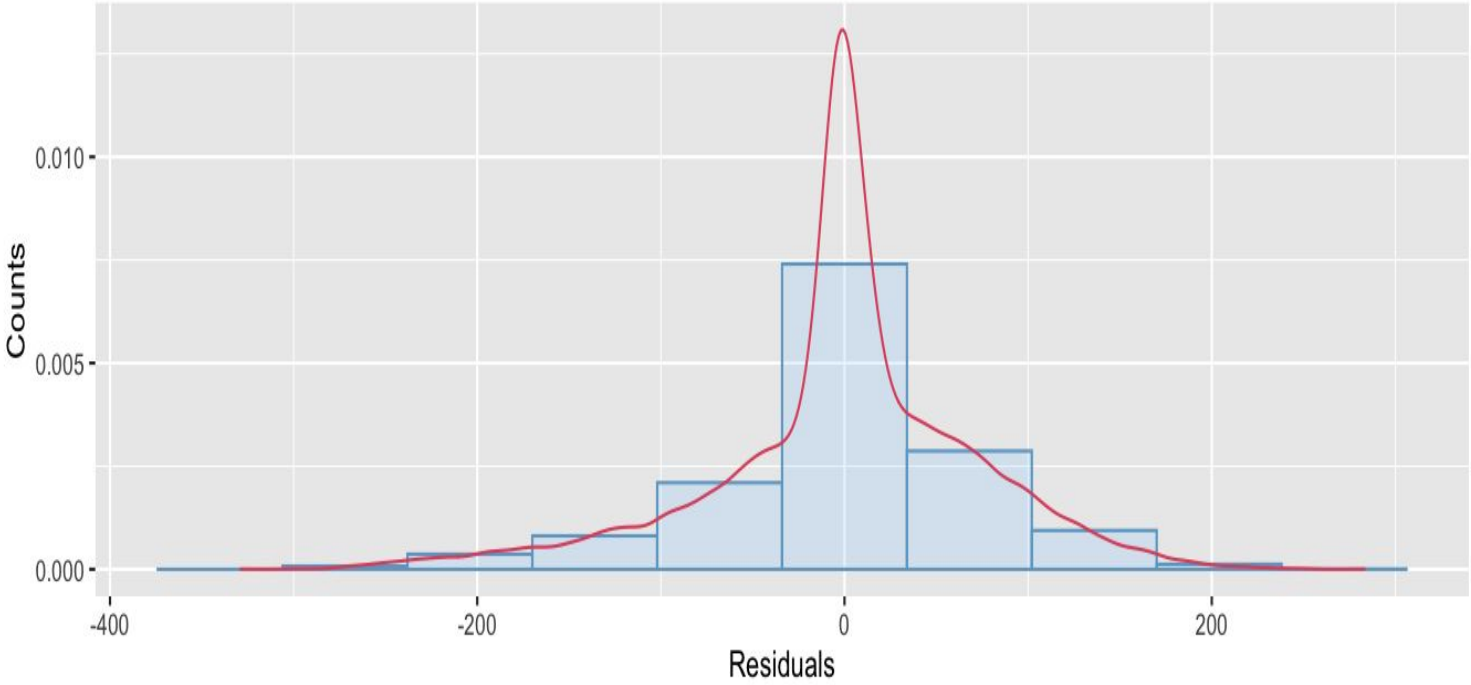
111.1554
Test RMSE

0.3154288
Train MAPE

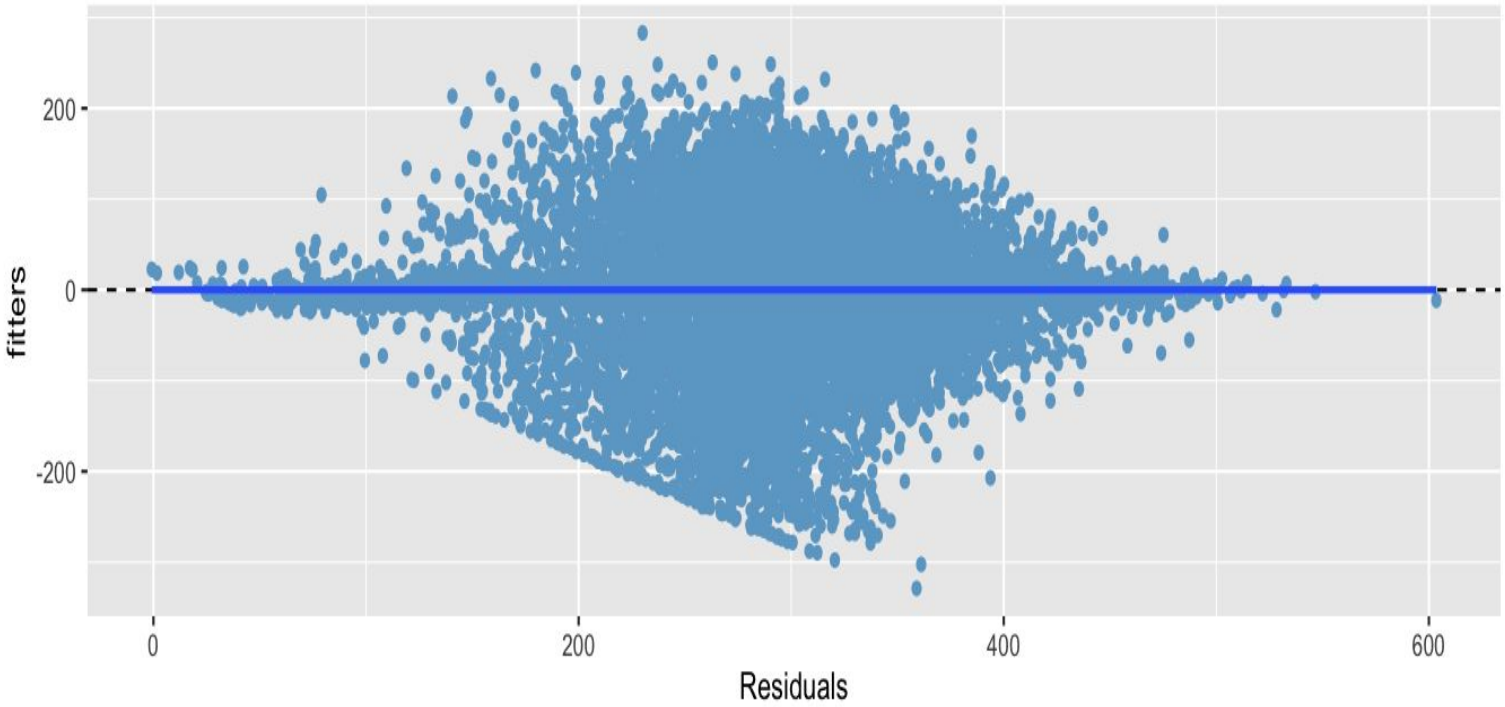
0.51242
Test MAPE

0.4405156
R Squared

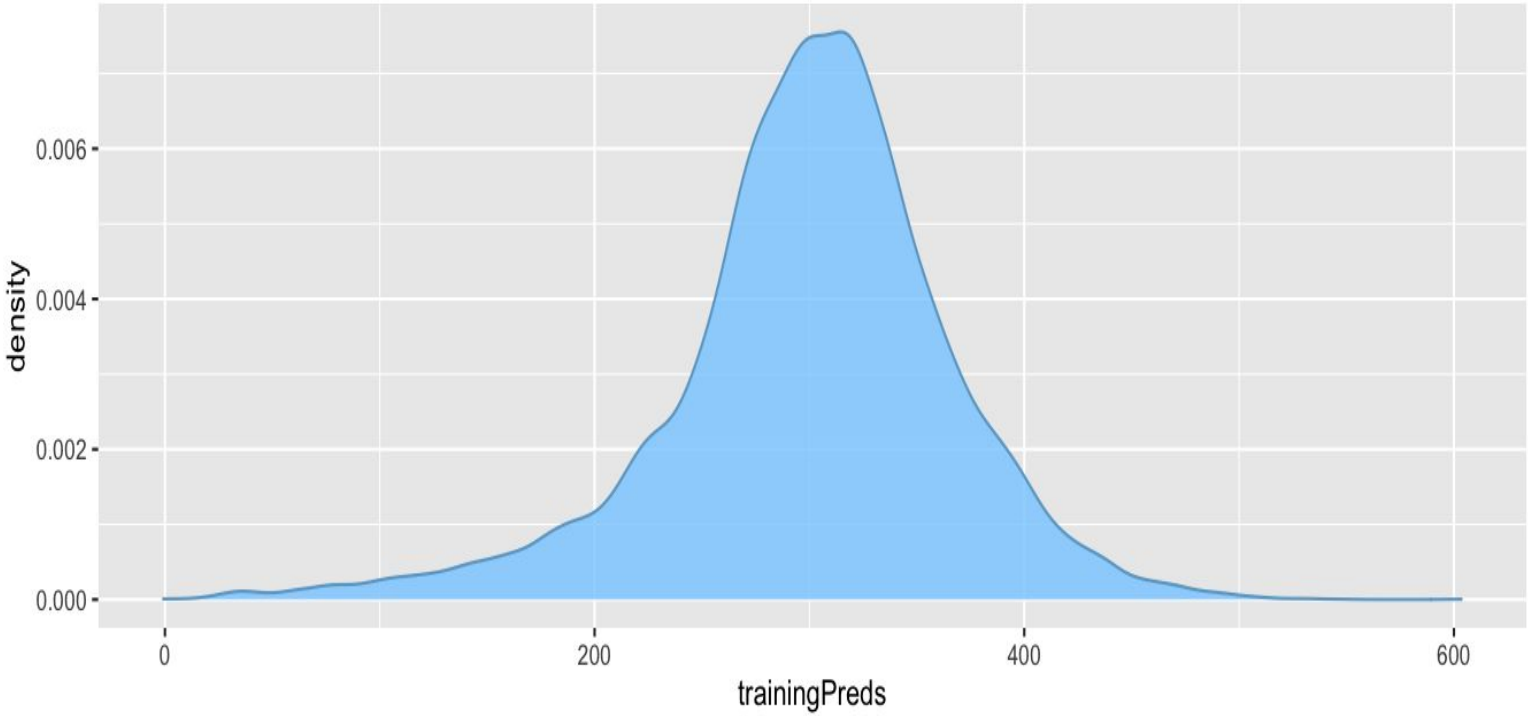
LR: Residual Histogram



LR: Residual vs Fitter



LR: Density training results



Random Forest

76.69371
Train RMSE

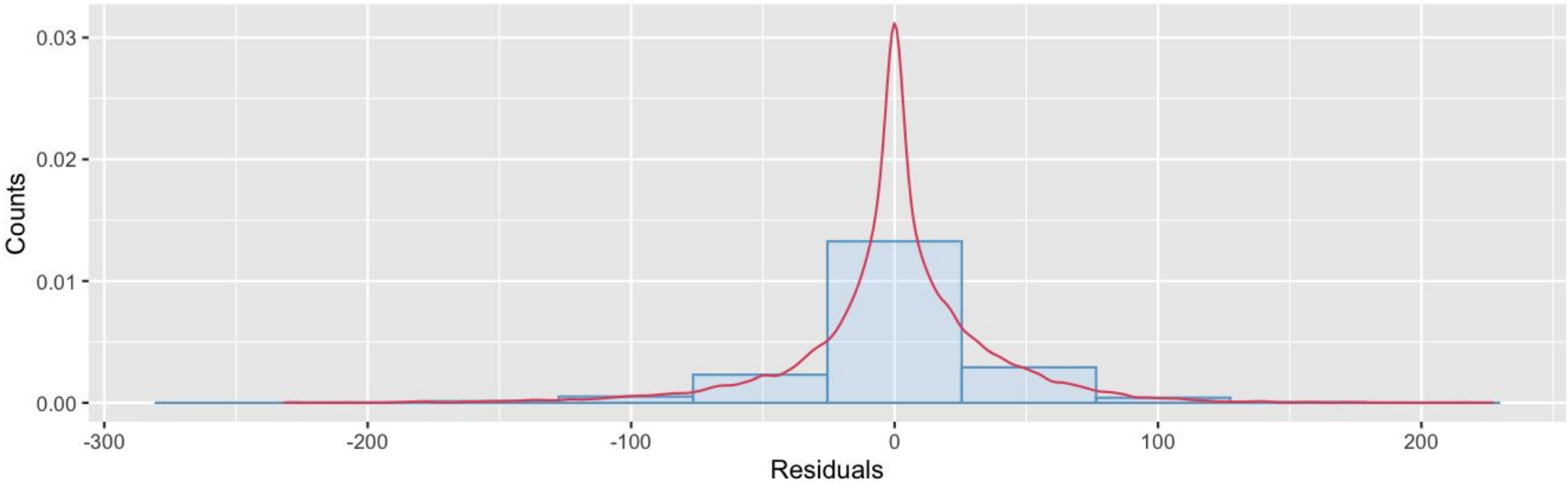
113.4756
Test RMSE

0.1433471
Train MAPE

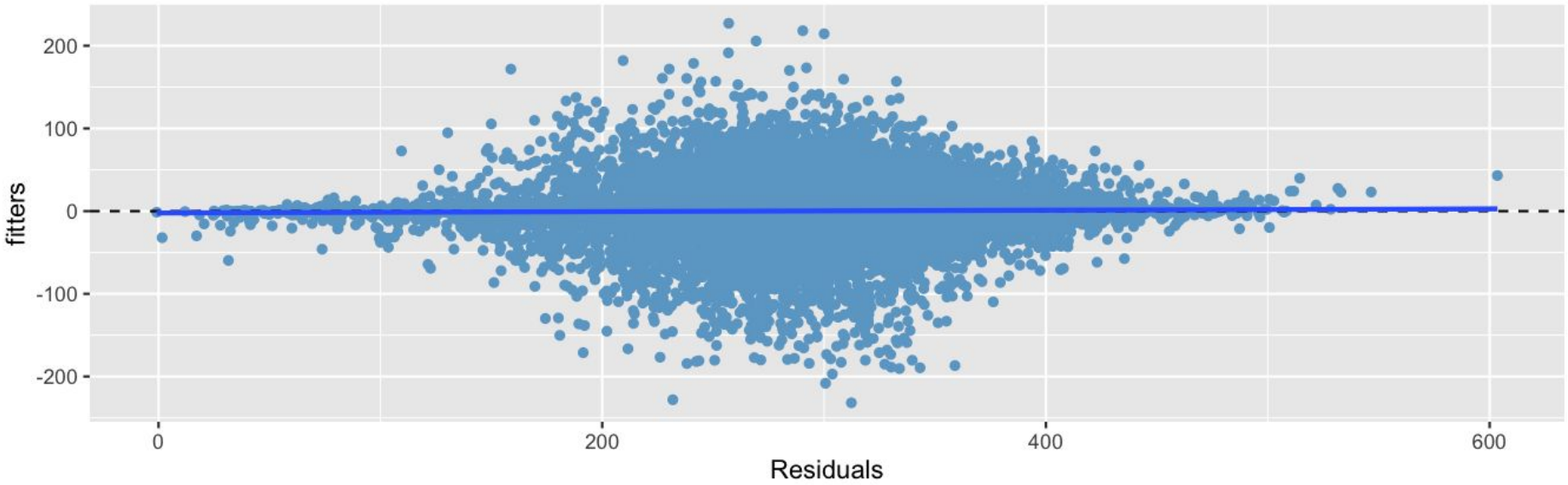
0.51242
Test MAPE

0.35982318
R Squared

RF: Residual Histogram



RF: Residual vs Fitter



RF: Training Density

k-Nearest Neighbors

85.59636
Train RMSE

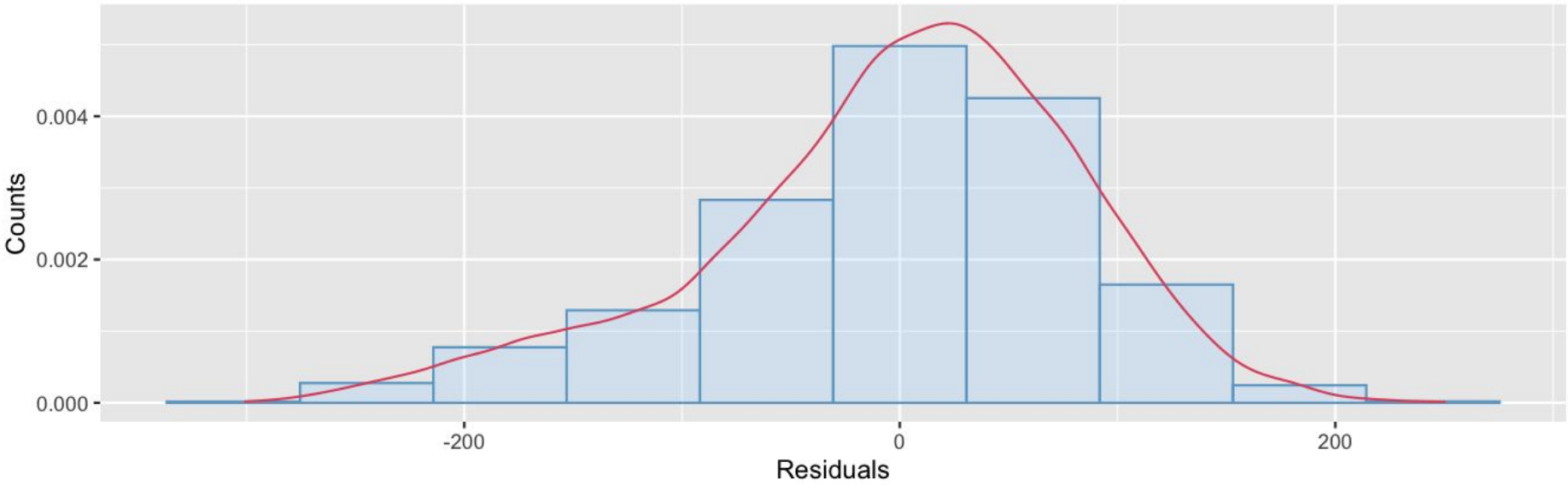
102.97
Test RMSE

0.4289104
Train MAPE

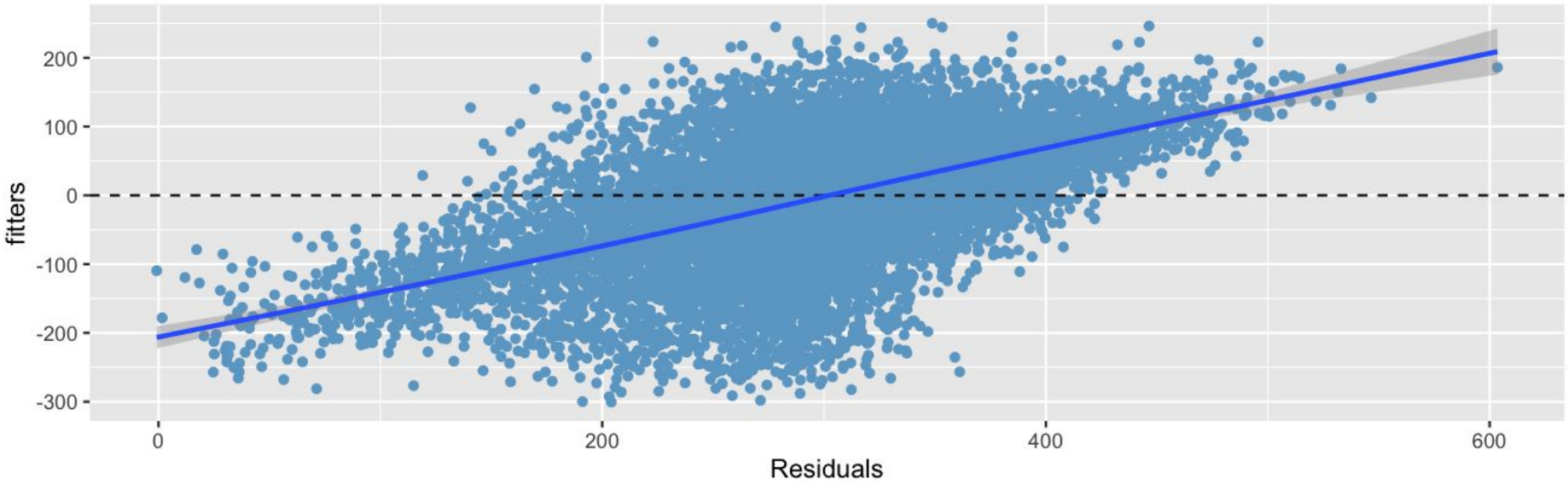
0.4961777
Test MAPE

0.03211571
R Squared

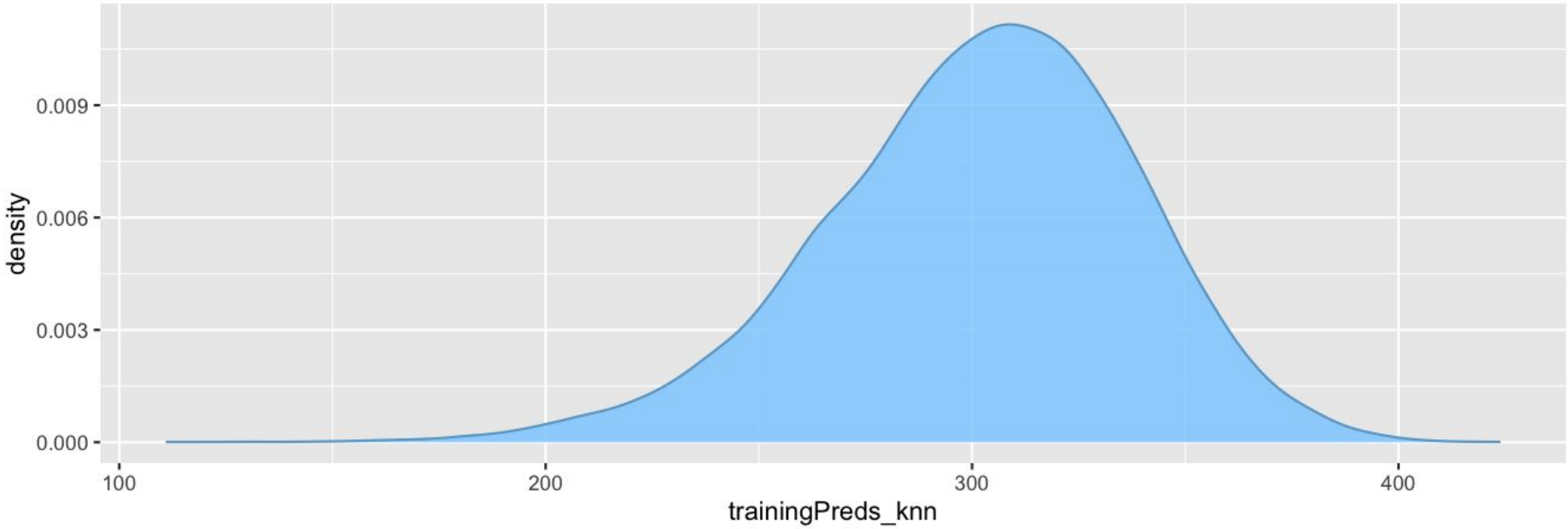
KNN: Residual Histogram



KNN: Residual vs Fitter



KNN: Training Density



04
Results

Model Compare

```
MAE
        Min.  1st Qu.   Median     Mean  3rd Qu.     Max. NA's
lm  50.70070 51.52673 51.90028 51.85798 52.28586 52.63721    0
rf  54.40517 55.25211 55.55989 55.57421 55.92425 56.86639    0
knn 77.65335 78.07331 78.31594 78.46230 78.70848 79.94792    0

RMSE
        Min.  1st Qu.    Median      Mean   3rd Qu.      Max. NA's
lm  72.50220 73.37546  73.90708  73.84974  74.36589  74.81487    0
rf  78.18727 79.44482  79.85081  79.81323  80.20784  81.36492    0
knn 99.15408 99.91990 100.29815 100.41223 100.76845 102.06685    0

Rsquared
          Min.   1st Qu.     Median       Mean    3rd Qu.       Max. NA's
lm  0.42737532 0.4362772 0.43903642 0.44051561 0.44443571 0.45616243    0
rf  0.33811871 0.3517198 0.36233822 0.35982318 0.36587561 0.37667955    0
knn 0.02982649 0.0338112 0.03697643 0.03716604 0.03896898 0.04693728    0
```

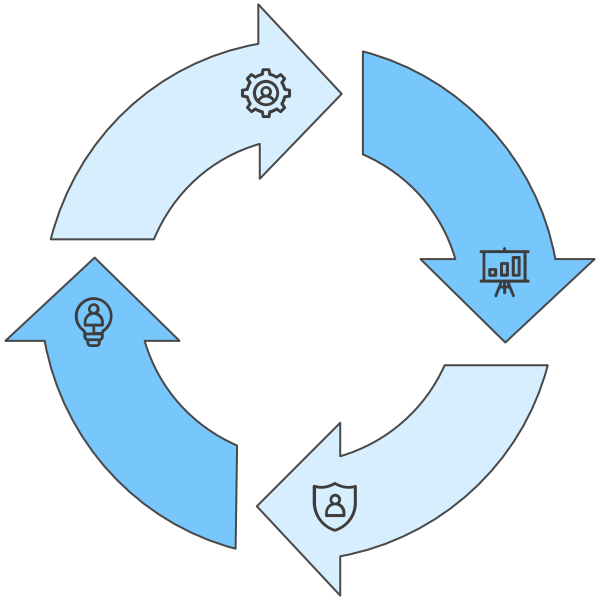
Linear Regression Model

KPI

Lowest RMSE
highest R^2

Residuals

The shape of the residuals symmetric and normally distributed



Residual vs Fitters

scatter plot of residuals are more or less randomly distributed around the 0 line

Hypothesis

We can assume LR Model can accurate predict the household spends in BBY

Sample of the predictions

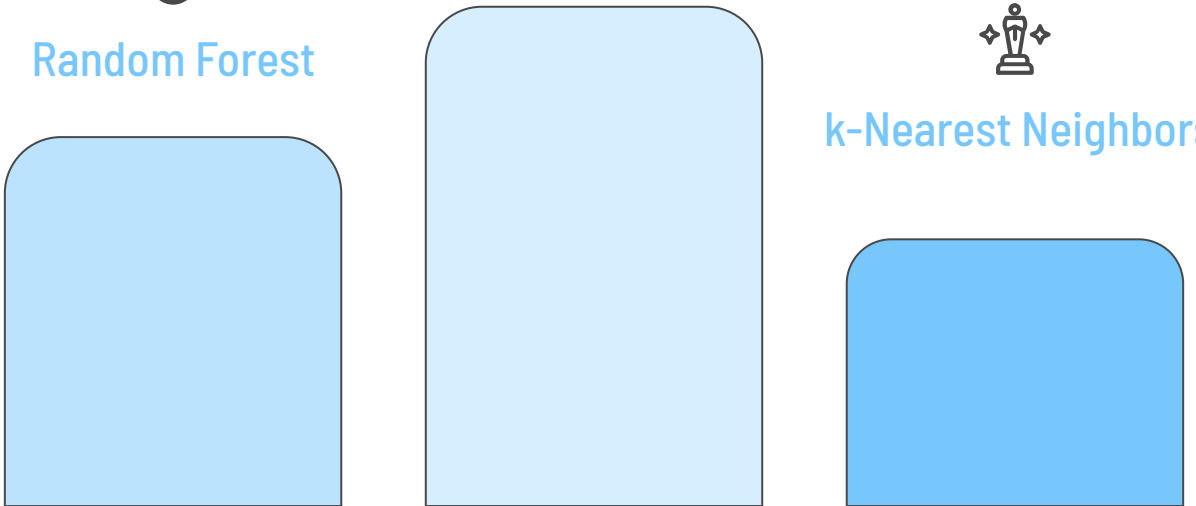
tmpID	FirsName	LastName	prospectSpend
1504	Horace	Goodwin	$249.65
328	Louise	Wisozk	$388.86
5448	Dorian	Douglas	$225.07
3907	Eden	Tremblay	$382.45
1728	Yer	King	$302.53
3716	Chong	Wilderman	$319.90
1920	Lamar	Dickinson	$299.72
5190	Roscoe	Grant	$296.54
4395	Wilburn	Collins	$303.44
1458	Virgil	Reichert	$300.65
4835	Marcos	Casper	$249.10
5943	Wally	Hegmann	$321.12
2090	Jeromy	Bashirian	$309.66
491	Hunter	Daugherty	$267.93
3909	Brunilda	Jewess	$275.01
51	Norberto	Lehner	$303.49
1732	Leeanne	Feeney	$274.75

Awards
Best Model

Linear Regression

Random Forest

k-Nearest Neighbors



Thanks!

Do you have any questions?

frf921@g.harvard.edu

References

"Accuracy Measures for a Forecast Model — Accuracy." n.d. Accessed May 9, 2022. https://pkg.robjhyndman.com/forecast/reference/accuracy.html.

"Correlation Test Between Two Variables in R - Easy Guides - Wiki - STHDA." n.d. Accessed May 9, 2022. http://www.sthda.com/english/wiki/correlation-test-between-two-variables-in-r.

"Ggplot2 Density Plot : Quick Start Guide - R Software and Data Visualization - Easy Guides - Wiki - STHDA." n.d. Accessed May 9, 2022. http://www.sthda.com/english/wiki/ggplot2-density-plot-quick-start-guide-r-software-and-data-visualization.

"Ggplot2 Quick Reference: Colour (and Fill) | Software and Programmer Efficiency Research Group." n.d. Accessed May 9, 2022. http://sape.inf.usi.ch/quick-reference/ggplot2/colour.

Holtz, Yan. n.d. "Basic Barplot with Ggplot2." Accessed May 9, 2022a. https://www.r-graph-gallery.com/218-basic-barplots-with-ggplot2.html.

———. n.d. "Basic Density Chart with Ggplot2." Accessed May 9, 2022b. https://www.r-graph-gallery.com/21-distribution-plot-using-ggplot2.html.

"How Do I Export Results from R Console? – QuickAdviser." n.d. Accessed May 9, 2022. https://quick-adviser.com/how-do-i-export-results-from-r-console/.

Johnson, Daniel. 2020. "R ANOVA Tutorial: One Way & Two Way (with Examples)." May 10, 2020. https://www.guru99.com/r-anova-tutorial.html.

"Kruskal-Wallis Test in R - Easy Guides - Wiki - STHDA." n.d. Accessed May 9, 2022. http://www.sthda.com/english/wiki/kruskal-wallis-test-in-r.

References

Mount, John. 2019. “What Is Vtreat? | R-Bloggers.” August 14, 2019. https://www.r-bloggers.com/2019/08/what-is-vtreat/.

“Plotting a Scatter Plot with Categorical Data. - General.” 2020. RStudio Community. June 11, 2020. https://community.rstudio.com/t/plotting-a-scatter-plot-with-categorical-data/69456.

“Predictive Modeling and Machine Learning in R with the Caret Package.” 2017. *Technical Tidbits From Spatial Analysis & Data Science* (blog). September 19, 2017. http://zevross.com/blog/2017/09/19/predictive-modeling-and-machine-learning-in-r-with-the-caret-package/.

“Random Forest Regression in R: Code and Interpretation | HackerNoon.” n.d. Accessed May 9, 2022. https://hackernoon.com/random-forest-regression-in-r-code-and-interpretation.

“RMSE: Root Mean Square Error.” n.d. Statistics How To. Accessed May 9, 2022. https://www.statisticshowto.com/probability-and-statistics/regression-analysis/rmse-root-mean-square-error/.

“RPubs - Correlation Coefficient Between Categorical and Continuous Variable.” n.d. Accessed May 9, 2022. https://rpubs.com/riazakhan94/correlation_between_categorical_and_continuous_variable.

Vandeput, Nicolas. 2021. “Forecast KPI: RMSE, MAE, MAPE & Bias.” Medium. July 30, 2021. https://towardsdatascience.com/forecast-kpi-rmse-mae-mape-bias-cdc5703d242d.

Wheeler, Willie. 2021. “Evaluating Linear Regression Models Using RMSE and R^2.” *Wwblog* (blog). June 24, 2021. https://medium.com/wwblog/evaluating-regression-models-using-rmse-and-r%C2%B2-42f77400efee.