HarvardX Data Science Adult Census Income Project

Chase2020 - HarvardX: PH125.9x

4/28/2020

Executive Summary

This report details the methods used to analyze the Adult Census Income Dataset. The reviewed dataset contains information from the 1994 Census Bureau Database. The original data obtained from the database was revised before it was uploaded to Kaggle.com for use. Some records were removed from the original data pull such as records for people under the age of 16 who worked less than 0 hours per week. The revised dataset from Kaggle.com was used for this project.

For each income record, the dataset included the following information: the person's age, working class, highest education level, education number corresponding to their education level, marital status, occupation, relationship, race, sex, capital gain reported, capital loss reported, hours worked per week, native country, identifier showing whether they made more or less than $50,000, and fnlwgt, which is a weighting metric used based on the person's demographic information. Based on this information, an additional data column was added to assist with this project. A data column showing the person's income level was added based on the existing income column. A one is shown if a person's annual income was less than or equal to $50,000, and a two is shown if their annual income was greater than $50,000. This column was added, so the dataset could be analyzed more efficiently.

The purpose of this review was to become familiar with the dataset and create two income prediction systems that can predict whether a person's income is greater than $50,000. To do this, the modified dataset was divided into two sections, which are labelled as training_v2 and validation_v2. Training_v2 was used to generate the models, and the validation_v2 dataset was used to test the income prediction systems. After utilizing data visualization techniques to analyze the data, the machine learning algorithms were developed in R using the Naïve Bayes and Random Forest models.

Data Preparation

The required packages were installed in R, and the dataset was imported from https://github.com/Rockefeller2020/Adult-Census-Income-Project-Dataset/raw/master/adult.csv. This is a copy of the original dataset downloaded from https://www.kaggle.com/uciml/adult-census-income. The original dataset was then modified, so the data could be analyzed more thoroughly. The income level variable was added. The dataset was then divided into two groups, training and validation. Ninety percent of the dataset was used in the training group to develop to algorithm, and the remaining ten percent was used in the validation group to test this. Using ninety percent of the dataset allows for more data to be used when building the algorithm, which should result in a more accurate model.

```
### Install required packages
if(!require(tidyverse)) install.packages("tidyverse", repos =
"http://cran.us.r-project.org")
if(!require(caret)) install.packages("caret", repos = "http://cran.us.r-
project.org")
if(!require(lubridate)) install.packages("lubridate", repos =
"http://cran.us.r-project.org")

library(randomForest)

### Import the dataset
# Dataset was downloaded originally from https://www.kaggle.com/uciml/adult-
census-income and then saved to a Github repository
dataset <- read.csv("https://github.com/Rockefeller2020/Adult-Census-Income-
Project-Dataset/raw/master/adult.csv")

### Modify the dataset to include an income_level item that can be used to
analyze the data
dataset_v2 <- dataset %>% mutate(income_level=as.numeric(income))

### Create the training and validation datasets. The validation dataset will
be 10% of the adult dataset.
set.seed(1)
test_index <- createDataPartition(y = dataset_v2$income, times=1, p=0.1, list
= FALSE)
training <- dataset_v2[-test_index,]
validation <- dataset_v2[test_index,]
```

Data Exploration

To gain a better understanding of the training dataset, a summary was pulled to show some high-level information. This shows the lowest income level is 1, and the highest is 2. This also shows the information that will be provided for each record along with the ranges or distributions within each.

```
### Methods and Data Analysis
### General analysis of the data
# Summary of the data
summary(training)

##       age                   workclass         fnlwgt
education
##  Min.   :17.00   Private         :20399   Min.   :  12285   HS-grad
:9449
##  1st Qu.:28.00   Self-emp-not-inc: 2285   1st Qu.: 117844   Some-
college:6550
##  Median :37.00   Local-gov       : 1890   Median : 178388   Bachelors
:4839
##  Mean   :38.52   ?               : 1662   Mean   : 189948   Masters
:1542
```

```
##  3rd Qu.:47.00   State-gov        : 1182   3rd Qu.: 237611   Assoc-voc
:1245
##  Max.   :90.00   Self-emp-inc     : 1002   Max.   :1455435   11th
:1066
##                  (Other)          :  884                     (Other)
:4613
##  education.num                marital.status           occupation
##  Min.   : 1.00   Divorced             : 3958   Prof-specialty :3724
##  1st Qu.: 9.00   Married-AF-spouse    :   20   Craft-repair   :3676
##  Median :10.00   Married-civ-spouse   :13461   Exec-managerial:3654
##  Mean   :10.08   Married-spouse-absent:  386   Adm-clerical   :3384
##  3rd Qu.:12.00   Never-married        : 9664   Sales          :3303
##  Max.   :16.00   Separated            :  925   Other-service  :2963
##                  Widowed              :  890   (Other)        :8600
##          relationship                   race            sex
##  Husband       :11873   Amer-Indian-Eskimo:  280   Female: 9690
##  Not-in-family : 7483   Asian-Pac-Islander:  915   Male  :19614
##  Other-relative:  892   Black             : 2813
##  Own-child     : 4564   Other             :  258
##  Unmarried     : 3092   White             :25038
##  Wife          : 1400
##
##   capital.gain    capital.loss      hours.per.week       native.country
##  Min.   :    0   Min.   :   0.00   Min.   : 1.00   United-States:26270
##  1st Qu.:    0   1st Qu.:   0.00   1st Qu.:40.00   Mexico       :  572
##  Median :    0   Median :   0.00   Median :40.00   ?            :  516
##  Mean   : 1073   Mean   :  86.98   Mean   :40.42   Philippines  :  173
##  3rd Qu.:    0   3rd Qu.:   0.00   3rd Qu.:45.00   Germany      :  120
##  Max.   :99999   Max.   :4356.00   Max.   :99.00   Canada       :  109
##                                                    (Other)      : 1544
##    income       income_level
##  <=50K:22248   Min.   :1.000
##  >50K : 7056   1st Qu.:1.000
##                Median :1.000
##                Mean   :1.241
##                3rd Qu.:1.000
##                Max.   :2.000
##
```

The first six rows of the dataset are shown below as examples. This shows the format of the information and is useful when determining how to analyze the data.

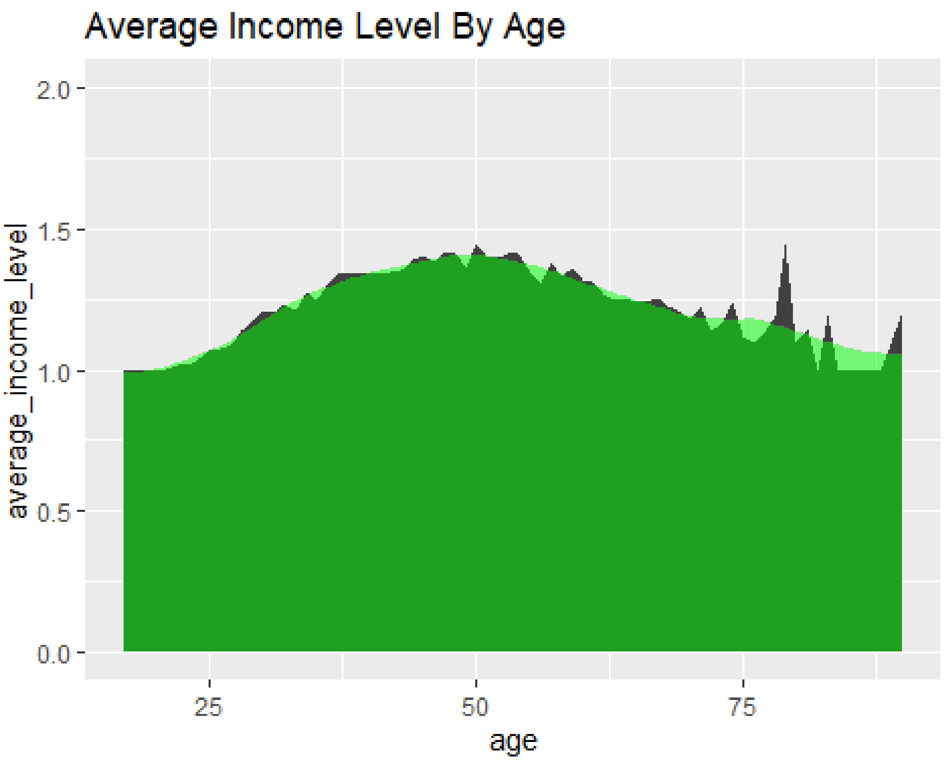
```
# Display the first six rows of the dataset
head(training)
```

```
##   age workclass fnlwgt    education education.num marital.status
## 1  90         ?  77053      HS-grad             9        Widowed
## 2  82   Private 132870      HS-grad             9        Widowed
## 3  66         ? 186061 Some-college            10        Widowed
## 4  54   Private 140359      7th-8th             4       Divorced
## 5  41   Private 264663 Some-college            10      Separated
## 6  34   Private 216864      HS-grad             9       Divorced
##          occupation  relationship  race    sex capital.gain capital.loss
## 1                 ? Not-in-family White Female            0         4356
## 2   Exec-managerial Not-in-family White Female            0         4356
## 3                 ?     Unmarried Black Female            0         4356
## 4 Machine-op-inspct     Unmarried White Female            0         3900
## 5    Prof-specialty     Own-child White Female            0         3900
## 6     Other-service     Unmarried White Female            0         3770
##   hours.per.week native.country income income_level
## 1             40  United-States  <=50K            1
## 2             18  United-States  <=50K            1
## 3             40  United-States  <=50K            1
## 4             40  United-States  <=50K            1
## 5             40  United-States  <=50K            1
## 6             45  United-States  <=50K            1
```

Data Analysis Through Visualization

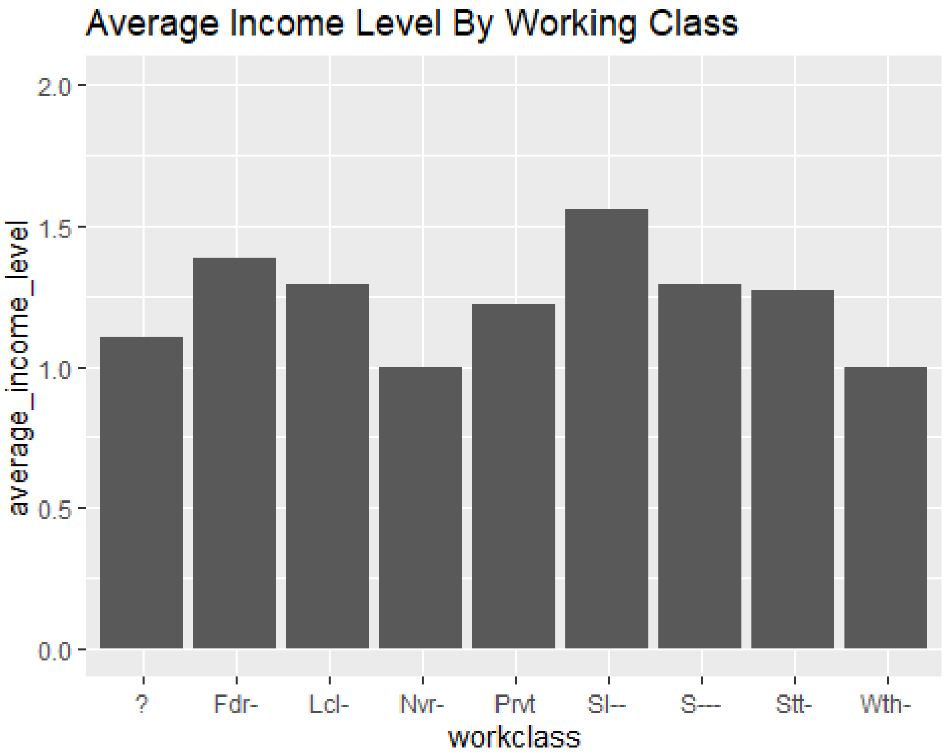
After analyzing the high-level information of the dataset, graphs were used to look deeper to identify trends that may be useful for the income prediction algorithms. These graphs use averages of some of the variables. The first graph shows the average income level by age. The income levels ranging from one to two were used, so the lowest average income level possible is one. The highest average income level possible is two. Averages in this graph show whether more people are earning more or less than $50,000 for the given age. For example, if the average income level is under 1.5 and closer to one, that means most people at that given age are earning less than or equal to $50,000. Using averages gives a better overall picture than looking at individual data points. This graph shows age affects a person’s income as more people at age 50 are earning more than $50,000 compared to those at age 25.

```
### Graphs showing average income levels for the dataset by variable
# Plot average income level by age
age_averages <- training %>%
  group_by(age) %>% summarize(average_income_level=mean(income_level))

ggplot(age_averages, aes(x = age, y = average_income_level)) +
  geom_area(stat = "identity", fill = "grey25") +
  ylim(0,2) +
  stat_smooth(geom = 'area', method = 'loess', span = 1/3,
              alpha = 1/2, fill = "green") +
  labs(title = "Average Income Level By Age")

## `geom_smooth()` using formula 'y ~ x'
```



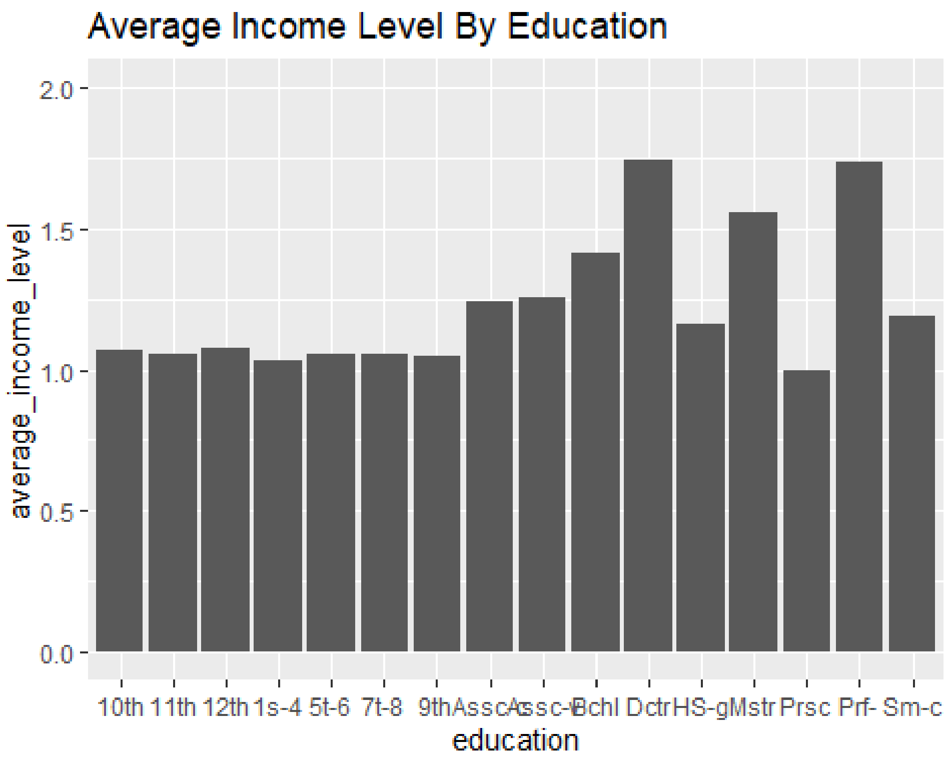
The next graph shows the average income levels by working class. This shows a person's working class will likely affect their annual income. Most people listed in the self-emp-inc (SI–) working class are making more than $50,000 annually, while most people in working classes such as the private (Prvt) and state-gov (Stt-) classes are making less than or equal to $50,000 annually.

```r
# Plot average income level by working class
workclass_averages <- training %>%
  group_by(workclass) %>% summarize(average_income_level=mean(income_level))

ggplot(workclass_averages, aes(x = workclass, y = average_income_level)) +
  geom_bar(stat = "identity") +
  scale_x_discrete(labels = abbreviate) +
  ylim(0,2) +
  labs(title = "Average Income Level By Working Class")
```



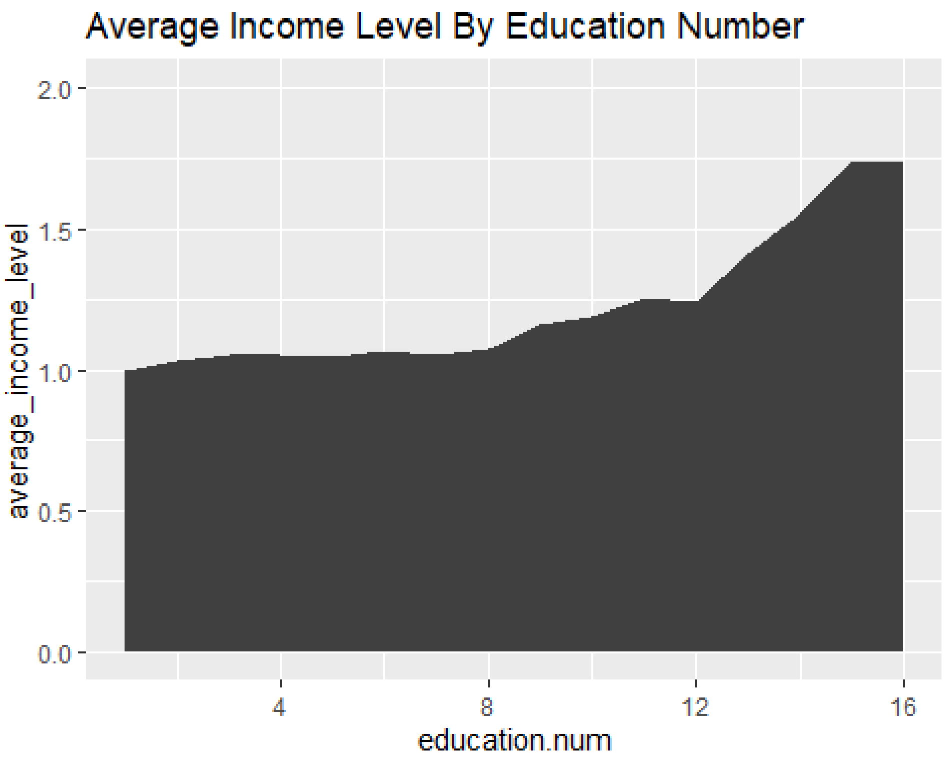
The graph below shows the average income levels by highest education level completed. This shows a person's education will likely impact whether they make more than $50,000 annually. The graph shows many who only have a high school degree or who never completed high school are making less than or equal to $50,000. The average income levels start rising the more education a person has completed.

```r
# Plot average income level by education
education_averages <- training %>%
  group_by(education) %>% summarize(average_income_level=mean(income_level))

ggplot(education_averages, aes(x = education, y = average_income_level)) +
  geom_bar(stat = "identity") +
  scale_x_discrete(labels = abbreviate) +
  ylim(0,2) +
  labs(title = "Average Income Level By Education")
```



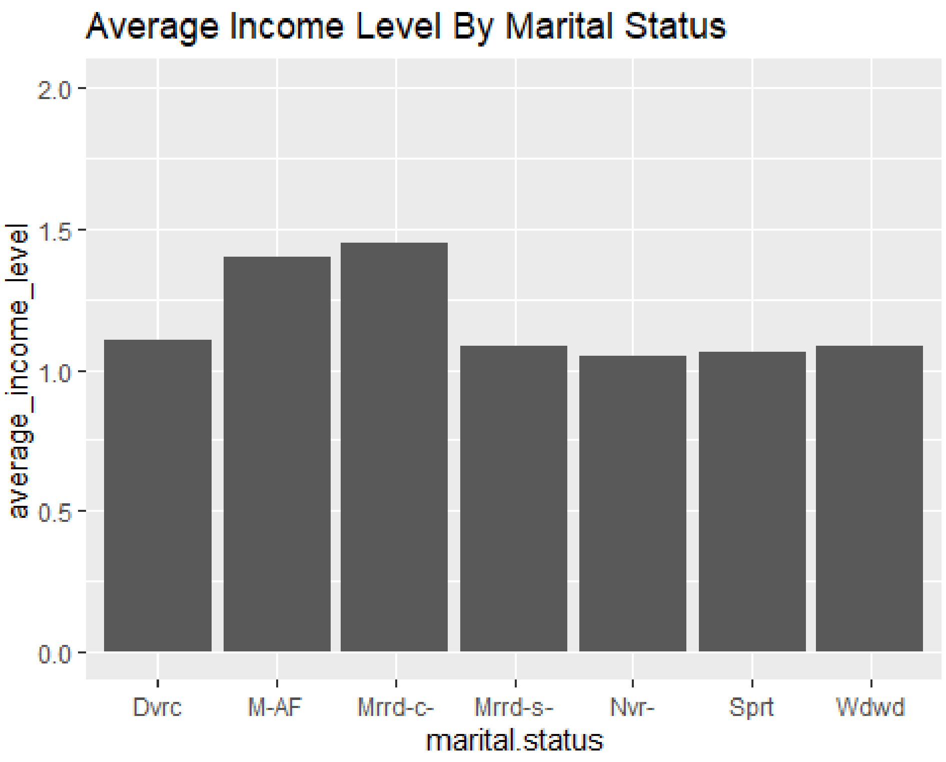
The next graph shows the average income levels by education number. This is very similar to the above graph. This simply assigns a number to the person's highest education earned on a scale of one to 16. For example, someone whose highest degree earned is a high school diploma will be assigned education number 9, and someone who has earned a bachelors degree will be assigned education number 13. This graph shows education has a large impact a person's income.

```
# Plot average income level by education number
education.num_averages <- training %>%
  group_by(education.num) %>%
summarize(average_income_level=mean(income_level))

ggplot(education.num_averages, aes(x = education.num, y =
average_income_level)) +
  geom_area(stat = "identity", fill = "grey25") +
  ylim(0,2) +
  labs(title = "Average Income Level By Education Number")
```



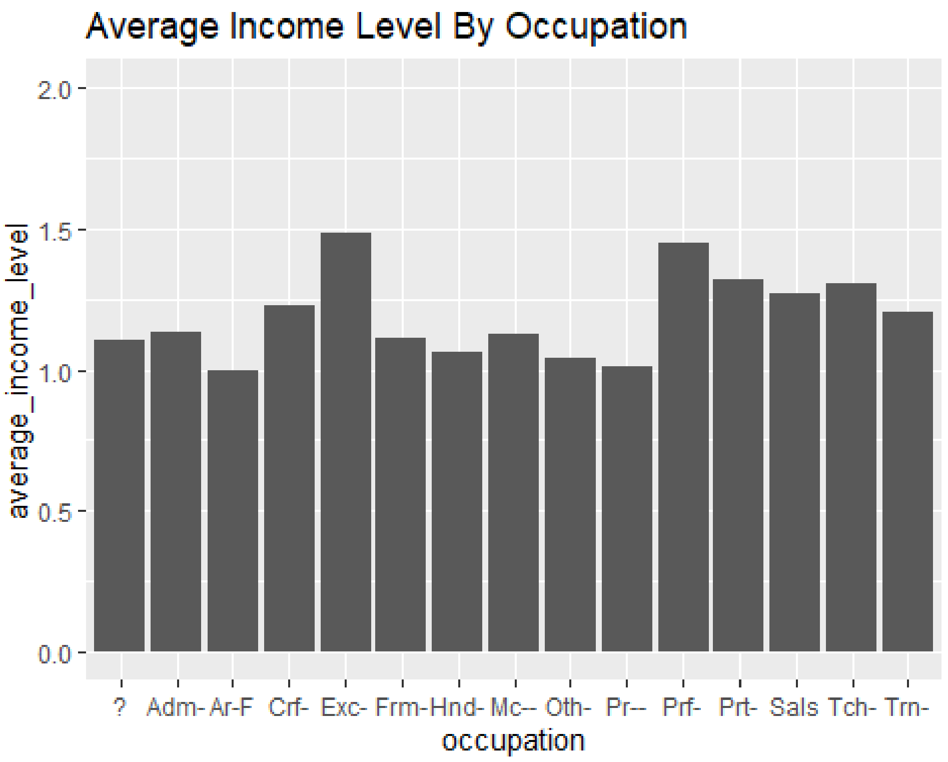
The following graph shows the average income levels by marital status. This shows a person's marital status may impact their annual income as more who are listed as married-AF-spouse (M-AF) and married-civ-spouse (Mrrd-c-) earn over $50,000 annually compared to the other statuses.

```r
# Plot average income level by marital status
marital.status_averages <- training %>%
  group_by(marital.status) %>%
summarize(average_income_level=mean(income_level))

ggplot(marital.status_averages, aes(x = marital.status, y =
average_income_level)) +
  geom_bar(stat = "identity") +
  scale_x_discrete(labels = abbreviate) +
  ylim(0,2) +
  labs(title = "Average Income Level By Marital Status")
```



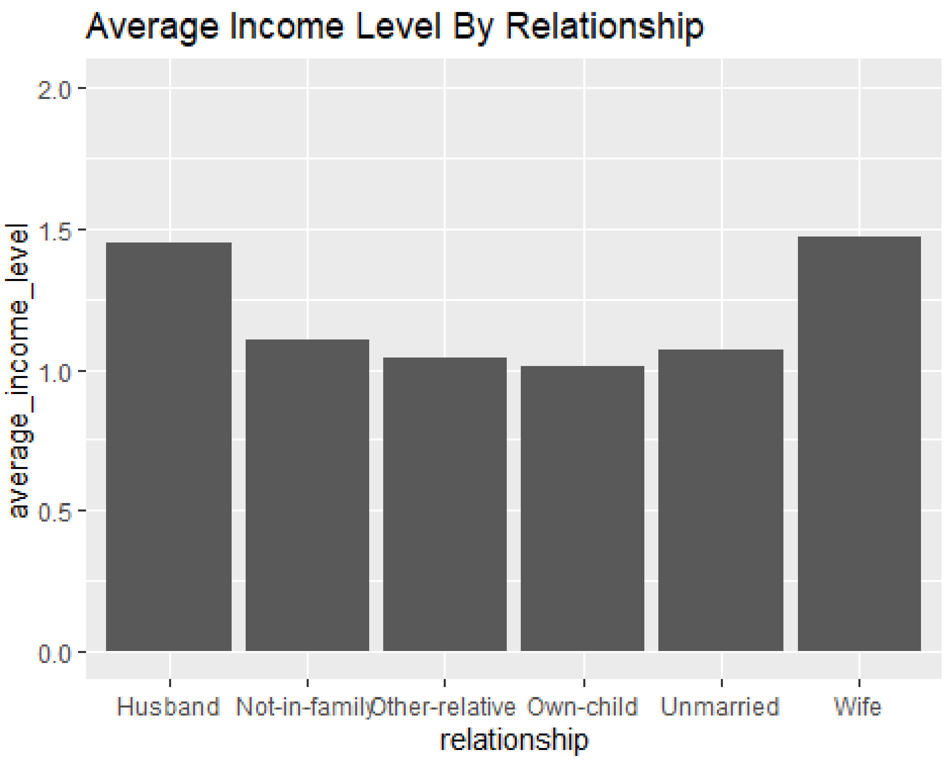
The next graph shows the average income levels by occupation. This shows a person's occupation will likely affect their annual income. For example, more people with the occupation exec-managerial (Exc-) are earning more than $50,000 annually compared to those with the occupation of adm-clerical (adm-).

```
# Plot average income level by occupation
occupation_averages <- training %>%
  group_by(occupation) %>% summarize(average_income_level=mean(income_level))

ggplot(occupation_averages, aes(x = occupation, y = average_income_level)) +
  geom_bar(stat = "identity") +
  scale_x_discrete(labels = abbreviate) +
  ylim(0,2) +
  labs(title = "Average Income Level By Occupation")
```



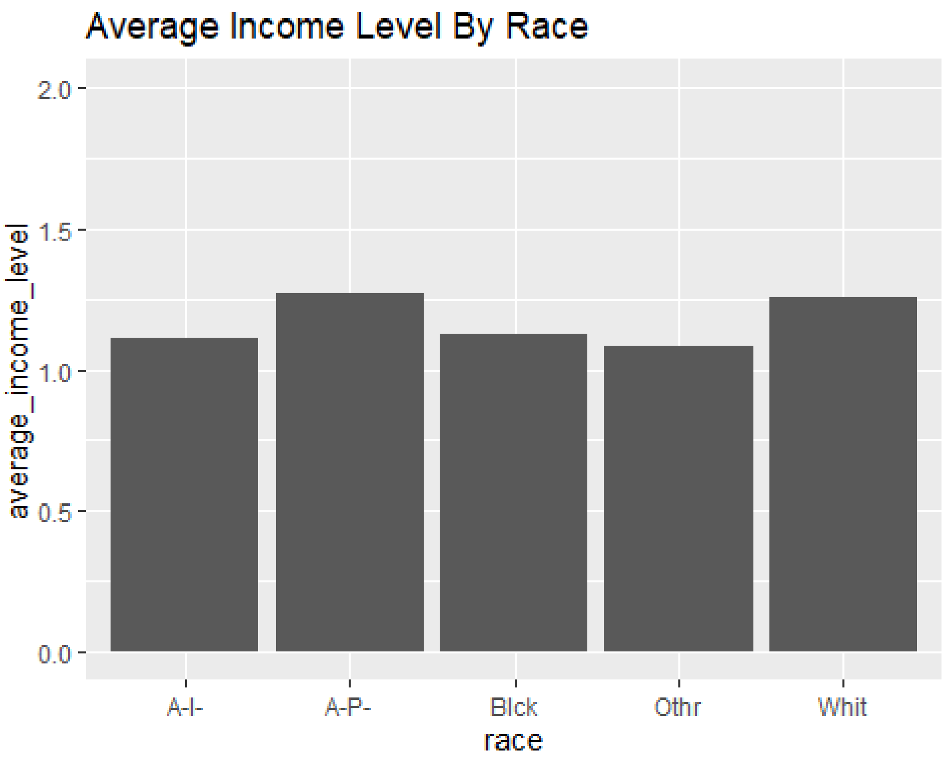
The below graph shows the average income levels by relationship. This shows a person's relationship may impact their annual income. More listed as husbands and wives earned over $50,000 annually than the other relationship categories.

```r
# Plot average income level by relationship
relationship_averages <- training %>%
  group_by(relationship) %>%
summarize(average_income_level=mean(income_level))

ggplot(relationship_averages, aes(x = relationship, y =
average_income_level)) +
  geom_bar(stat = "identity") +
  ylim(0,2) +
  labs(title = "Average Income Level By Relationship")
```



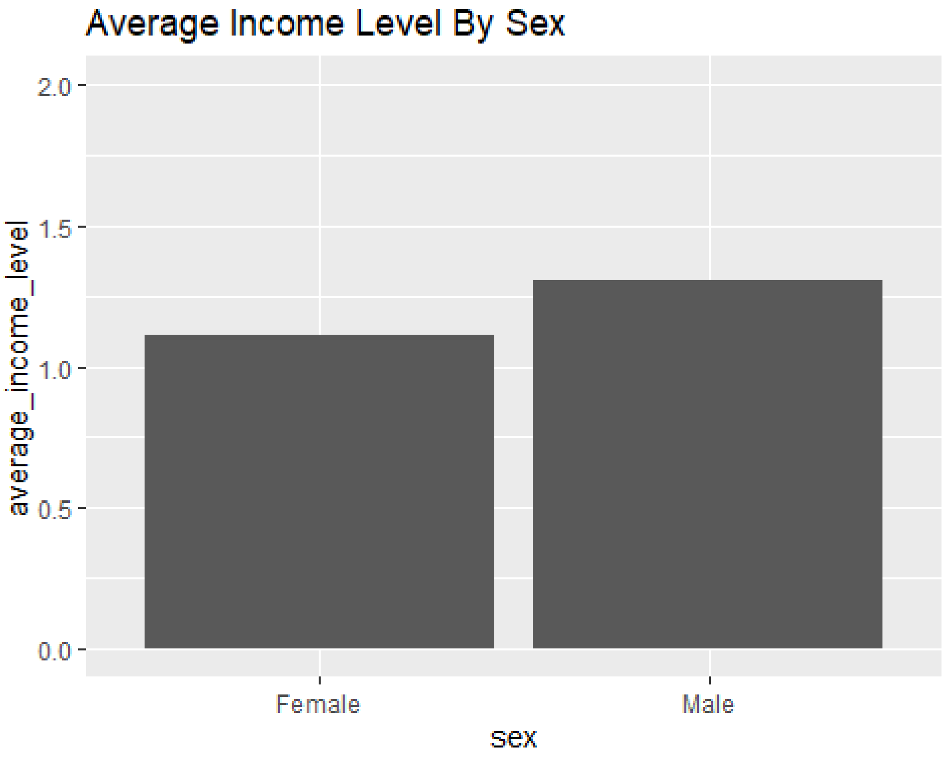
The next graph shows the average income levels by race. This shows a person's race may impact their annual salary. The asian-pac-islander (A-P-) and white (Whit) races on average have more people earning more than $50,000 annually than the other races.

```r
# Plot average income level by race
race_averages <- training %>%
  group_by(race) %>% summarize(average_income_level=mean(income_level))

ggplot(race_averages, aes(x = race, y = average_income_level)) +
  geom_bar(stat = "identity") +
  scale_x_discrete(labels = abbreviate) +
  ylim(0,2) +
  labs(title = "Average Income Level By Race")
```



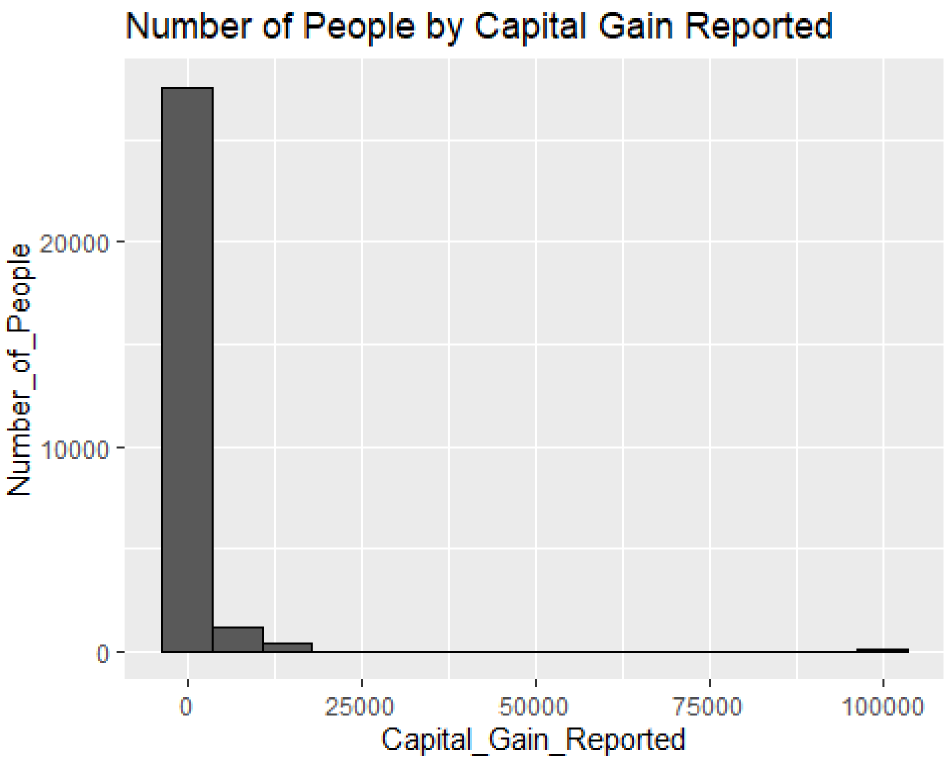
The following graph shows the average income levels by sex. Based on this graph, it appears there are more males earning more than $50,000 annually than females.

```
# Plot average income level by sex
sex_averages <- training %>%
  group_by(sex) %>% summarize(average_income_level=mean(income_level))

ggplot(sex_averages, aes(x = sex, y = average_income_level)) +
  geom_bar(stat = "identity") +
  ylim(0,2) +
  labs(title = "Average Income Level By Sex")
```



Based on the summary information, it appeared that many records in the dataset showed $0 for the capital gain reported. The below graph shows the number of people reporting a given capital gain. The graph confirms most did not report a capital gain. An exact count of the number of people reporting a capital gain of $0 was pulled showing over 90% of the dataset shows $0 for capital gains reported. As this applies to a very small portion of the population, it may not be a good indicator of a person's annual income.

```
# Plot average income level by capital gain reported
training %>%
  ggplot(aes(capital.gain)) +
  geom_histogram(bins = 15, color = "black") +
  labs(title = "Number of People by Capital Gain Reported",
       x = "Capital_Gain_Reported",
       y = "Number_of_People")
```

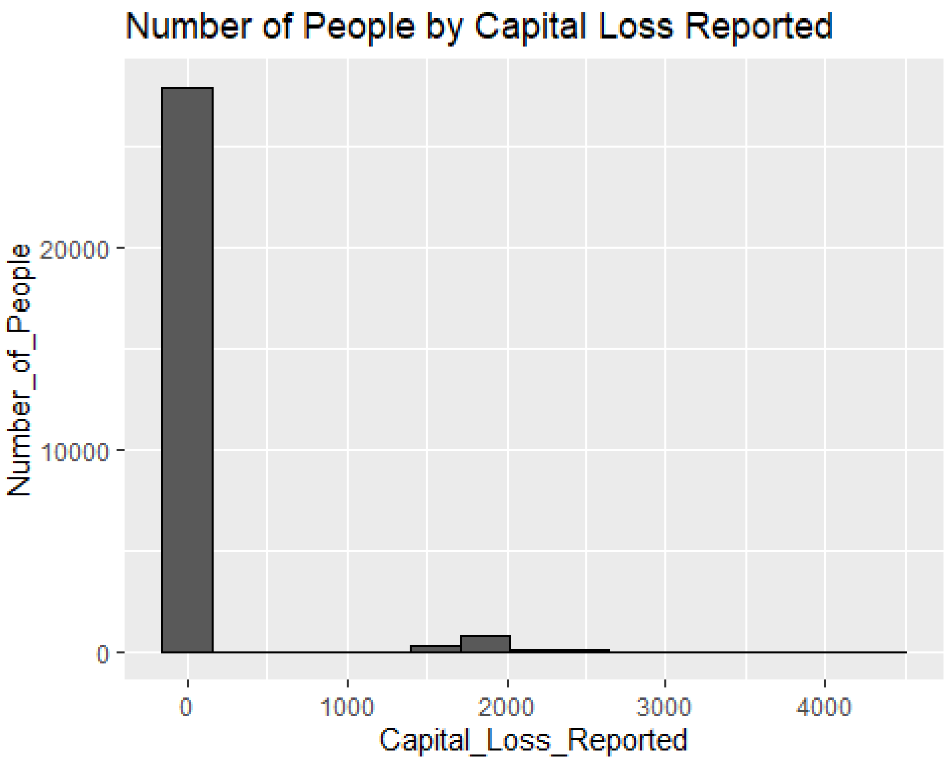


```
# Determine number of people reporting a captial gain of $0
training %>%
  count(capital.gain == 0)
```

```
## # A tibble: 2 x 2
##   `capital.gain == 0`     n
##   <lgl>               <int>
## 1 FALSE                2447
## 2 TRUE                26857
```

Based on the summary information, it also appeared that many records in the dataset showed $0 for the capital loss reported. The below graph shows the number of people reporting a given capital loss. The graph confirms most did not report a capital loss. An exact count of the number of people reporting a capital loss of $0 was pulled showing over 90% of the dataset shows $0 for capital loss reported. As this applies to a very small portion of the population, it also may not be a good indicator of a person's annual income.

```
# Plot average income level by capital loss reported
training %>%
  ggplot(aes(capital.loss)) +
  geom_histogram(bins = 15, color = "black") +
  labs(title = "Number of People by Capital Loss Reported",
       x = "Capital_Loss_Reported",
       y = "Number_of_People")
```

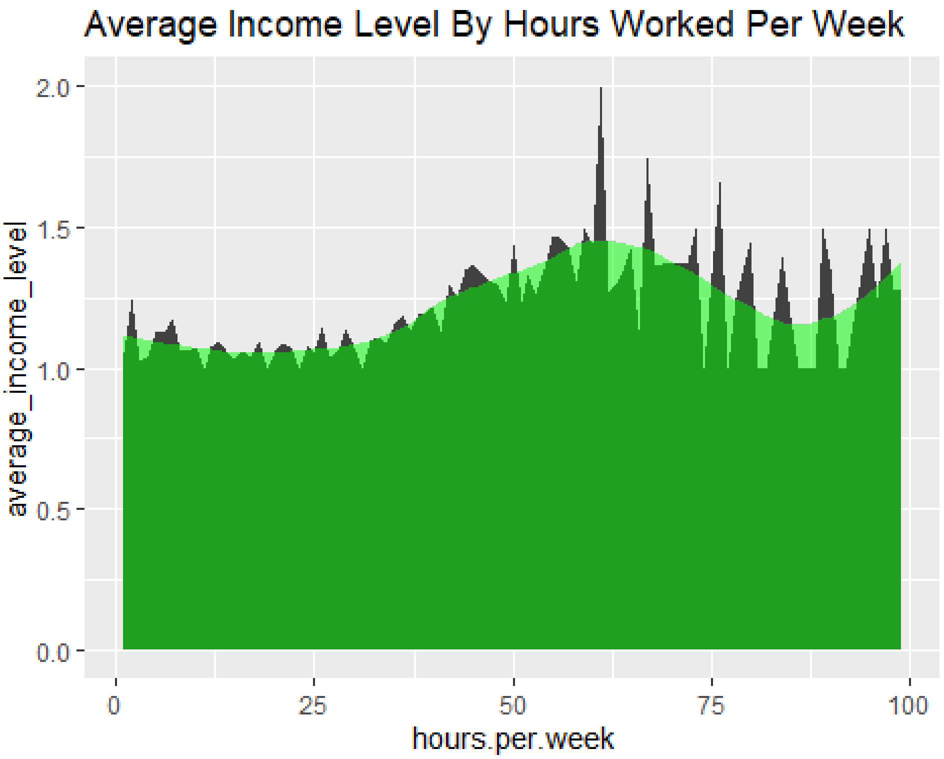


```
# Determine number of people reporting a captial loss of $0
training %>%
  count(capital.loss == 0)

## # A tibble: 2 x 2
##   `capital.loss == 0`     n
##   <lgl>                <int>
## 1 FALSE                 1358
## 2 TRUE                 27946
```

The next graph shows the average income levels by hours worked per week. This graph is interesting as it shows working more hours does not necessarily mean a person will make more. This shows people working approximately 60 hours per week are the most likely to be making more than $50,000 annually.

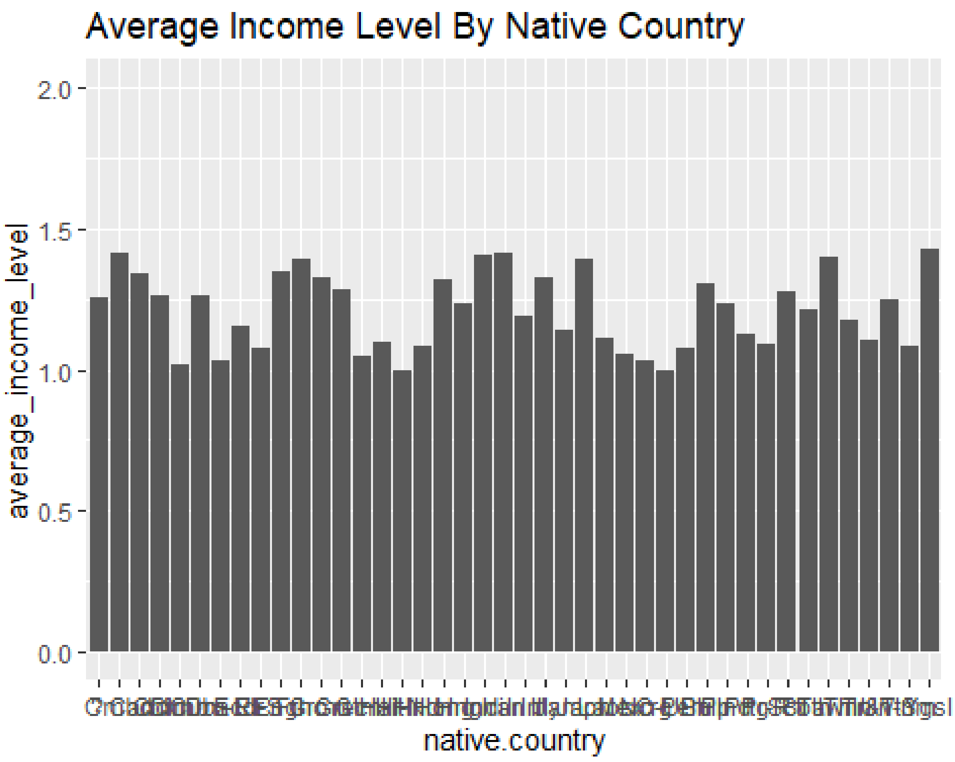
```
# Plot average income level by hours worked per week
hours.per.week_averages <- training %>%
  group_by(hours.per.week) %>%
summarize(average_income_level=mean(income_level))

ggplot(hours.per.week_averages, aes(x = hours.per.week, y =
average_income_level)) +
  geom_area(stat = "identity", fill = "grey25") +
  ylim(0,2) +
  stat_smooth(geom = 'area', method = 'loess', span = 1/3,
              alpha = 1/2, fill = "green") +
  labs(title = "Average Income Level By Hours Worked Per Week")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



The last graph shows the average income levels by native country. This graph shows on average more people from certain countries make more than $50,000 annually compared to other countries.

```r
# Plot average income level by native country
native.country_averages <- training %>%
  group_by(native.country) %>%
summarize(average_income_level=mean(income_level))

ggplot(native.country_averages, aes(x = native.country, y =
average_income_level)) +
  geom_bar(stat = "identity") +
  scale_x_discrete(labels = abbreviate) +
  ylim(0,2) +
  labs(title = "Average Income Level By Native Country")
```



Income Prediction Models

Based on the trends and averages identified through the analysis of the dataset, the income prediction models can begin to be built. The first step taken was to remove certain variables from the training and validation datasets that will not be used in the models. The fnlwgt variable was removed as this is used to weight records based on demographic information, but it is not necessarily useful when determining a person's income level. The capital gain and capital loss variables were removed as a large portion of the dataset showed capital gains and losses of $0. It was determined these may not be the best indicators of a person's income level either. Finally, the education and income level variables were removed as these are essentially duplicates of the education number and income variables.

```
### Models
# Remove variables not needed by models in both training and validation
datasets
training_v2 <- within(training, rm(fnlwgt, capital.gain, capital.loss,
education, income_level))
summary(training_v2)

##       age                   workclass     education.num
##  Min.   :17.00   Private         :20399   Min.   : 1.00
##  1st Qu.:28.00   Self-emp-not-inc: 2285   1st Qu.: 9.00
##  Median :37.00   Local-gov       : 1890   Median :10.00
##  Mean   :38.52   ?               : 1662   Mean   :10.08
##  3rd Qu.:47.00   State-gov       : 1182   3rd Qu.:12.00
##  Max.   :90.00   Self-emp-inc    : 1002   Max.   :16.00
##                   (Other)         :  884
##                 marital.status             occupation           relationship
##  Divorced             : 3958   Prof-specialty :3724   Husband       :11873
##  Married-AF-spouse    :   20   Craft-repair   :3676   Not-in-family : 7483
##  Married-civ-spouse   :13461   Exec-managerial:3654   Other-relative:  892
##  Married-spouse-absent:  386   Adm-clerical   :3384   Own-child     : 4564
##  Never-married        : 9664   Sales          :3303   Unmarried     : 3092
##  Separated            :  925   Other-service  :2963   Wife          : 1400
##  Widowed              :  890   (Other)        :8600
##                   race           sex        hours.per.week
##  Amer-Indian-Eskimo:  280   Female: 9690   Min.   : 1.00
##  Asian-Pac-Islander:  915   Male  :19614   1st Qu.:40.00
##  Black             : 2813                  Median :40.00
##  Other             :  258                  Mean   :40.42
##  White             :25038                  3rd Qu.:45.00
##                                            Max.   :99.00
##
##        native.country    income
##  United-States:26270   <=50K:22248
##  Mexico       :  572   >50K : 7056
##  ?            :  516
##  Philippines  :  173
##  Germany      :  120
##  Canada       :  109
##  (Other)      : 1544
```

```
validation_v2 <- within(validation, rm(fnlwgt, capital.gain, capital.loss,
education, income_level))
summary(validation_v2)
```

```
##       age                   workclass    education.num
##  Min.   :17.00   Private         :2297   Min.   : 1.00
##  1st Qu.:28.00   Self-emp-not-inc: 256   1st Qu.: 9.00
##  Median :38.00   Local-gov       : 203   Median :10.00
##  Mean   :39.17   ?               : 174   Mean   :10.07
##  3rd Qu.:49.00   State-gov       : 116   3rd Qu.:12.00
##  Max.   :90.00   Self-emp-inc    : 114   Max.   :16.00
##                  (Other)         :  97
##                 marital.status           occupation          relationship
##  Divorced             : 485   Craft-repair   :423   Husband       :1320
##  Married-AF-spouse    :   3   Prof-specialty :416   Not-in-family : 822
##  Married-civ-spouse   :1515   Exec-managerial:412   Other-relative:  89
##  Married-spouse-absent:  32   Adm-clerical   :386   Own-child     : 504
##  Never-married        :1019   Sales          :347   Unmarried     : 354
##  Separated            : 100   Other-service  :332   Wife          : 168
##  Widowed              : 103   (Other)        :941
##                   race          sex       hours.per.week
native.country
##  Amer-Indian-Eskimo:  31   Female:1081   Min.   : 1.00   United-
```

```
States:2900
##  Asian-Pac-Islander: 124   Male  :2176   1st Qu.:40.00   Mexico       :
71
##  Black             : 311                 Median :40.00   ?            :
67
##  Other             :  13                 Mean   :40.63   Philippines  :
25
##  White             :2778                 3rd Qu.:45.00   Germany      :
17
##                                          Max.   :99.00   Puerto-Rico  :
14
##                                                          (Other)      :
163
##    income
##  <=50K:2472
##  >50K : 785
##
##
##
##
##
```

The first model built is based on the Naïve Bayes Model. This model uses input attributes to determine a specific output. The naïve assumption is that all attributes are independent. The training_v2 dataset was separated into input attributes and prediction possibilities. From there, the model is built and used to predict people's incomes using the validation_v2 dataset. The confusion matrix results are shown displaying the accuracy of this model. The details of the confusion matrix results will be discussed in the results section of this report.

```
# Naïve Bayes Model
# Separate the income variable being predicted from the other variables in
the dataset
training_attributes <- training_v2[,-11]
training_income <- training_v2$income
# Build the naïve bayes model
naïve_bayes_model <- train(training_attributes, training_income, 'nb',
trControl = trainControl(method = 'cv', number = 10))
# Use the naïve bayes model to predict the income variable for the
validation_v2 dataset
naïve_bayes_prediction <- predict(naïve_bayes_model, validation_v2)
# Display confusion matrix results for the naïve bayes model
naive_bayes_cm <- confusionMatrix(data = naïve_bayes_prediction, reference =
validation_v2$income, positive = '>50K')
naive_bayes_cm

## Confusion Matrix and Statistics
##
##             Reference
## Prediction <=50K >50K
##      <=50K  2302  385
##      >50K    170  400
##
##                Accuracy : 0.8296
##                  95% CI : (0.8162, 0.8424)
##     No Information Rate : 0.759
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.4862
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.5096
##             Specificity : 0.9312
##          Pos Pred Value : 0.7018
##          Neg Pred Value : 0.8567
##              Prevalence : 0.2410
##          Detection Rate : 0.1228
##    Detection Prevalence : 0.1750
##       Balanced Accuracy : 0.7204
##
##        'Positive' Class : >50K
##
```

The second model is based on the Random Forest Model. This model uses numerous different decision trees to simulate income predictions based on the other attributes in the dataset. The most popular prediction from the decision tree simulations will be used as the prediction. The model was built using the training_v2 dataset and used to predict people's incomes in the validation_v2 dataset. The confusion matrix results will be discussed in the results section.

```
# Random Forest Model
# Build the random forest model
random_forest_model <- randomForest(income ~ ., data = training_v2, ntree =
2500)
# Use the random forest model to predict the income variable for the
validation_v2 dataset
random_forest_prediction <- predict(random_forest_model, validation_v2)
# Display confusion matrix results for the random forest model
random_forest_cm <- confusionMatrix(data = random_forest_prediction,
reference = validation_v2$income, positive = '>50K')
random_forest_cm
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction <=50K >50K
##      <=50K  2241  311
##      >50K    231  474
##
##                Accuracy : 0.8336
##                  95% CI : (0.8203, 0.8462)
##     No Information Rate : 0.759
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.5288
##
##  Mcnemar's Test P-Value : 0.0006905
##
##             Sensitivity : 0.6038
##             Specificity : 0.9066
##          Pos Pred Value : 0.6723
##          Neg Pred Value : 0.8781
##              Prevalence : 0.2410
##          Detection Rate : 0.1455
##    Detection Prevalence : 0.2165
##       Balanced Accuracy : 0.7552
##
##        'Positive' Class : >50K
##
```

Results

The confusion matrix results for both models are shown below. Overall accuracy, sensitivity, and specificity are the indictors that will be used to determine how well the models performed overall. Overall accuracy shows the percentage of incomes greater than or less than or equal to $50,000 the model was able to accurately predict for the validation_v2 dataset. Sensitivity shows the percentage where the model predicted greater than $50,000 for records where the person's income was greater than $50,000. Specificity shows the percentage where the model predicted less than or equal to $50,000 where the person's income was less than or equal to $50,000.

Regarding sensitivity, it appears the Random Forest Model was more accurate by a margin of approximately 9%. The Naïve Bayes Model was more accurate when comparing for specificity. The Naïve Bayes Model had a specificity of 2% higher than the Random Forest Model. The Random Forest Model had a slightly higher overall accuracy score of 83.36% compared to the Naïve Bayes Model's accuracy score of 82.96%. Overall, the Random Forest Model was able to more accurately predict incomes in the validation_v2 dataset.

```
### Results
# Display confusion matrix results for both models
naive_bayes_cm
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction <=50K >50K
##      <=50K  2302  385
##      >50K    170  400
##
##                Accuracy : 0.8296
##                  95% CI : (0.8162, 0.8424)
##     No Information Rate : 0.759
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.4862
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.5096
##             Specificity : 0.9312
##          Pos Pred Value : 0.7018
##          Neg Pred Value : 0.8567
##              Prevalence : 0.2410
##          Detection Rate : 0.1228
##    Detection Prevalence : 0.1750
##       Balanced Accuracy : 0.7204
##
##        'Positive' Class : >50K
##
```

```
random_forest_cm
```

```
## Confusion Matrix and Statistics
##
##            Reference
## Prediction <=50K >50K
##      <=50K  2241  311
##      >50K    231  474
##
##                Accuracy : 0.8336
##                  95% CI : (0.8203, 0.8462)
##     No Information Rate : 0.759
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.5288
##
##  Mcnemar's Test P-Value : 0.0006905
##
##             Sensitivity : 0.6038
##             Specificity : 0.9066
##          Pos Pred Value : 0.6723
##          Neg Pred Value : 0.8781
##              Prevalence : 0.2410
##          Detection Rate : 0.1455
##    Detection Prevalence : 0.2165
##       Balanced Accuracy : 0.7552
##
##        'Positive' Class : >50K
##
```

Conclusion

In conclusion, both the Naïve Bayes and Random Forest Models were able to predict incomes in the validation_v2 dataset. The accuracy of the predictions ranged between 82.96% and 83.36%. This is considered to be reasonably accurate based on the given dataset. A person's income can range greatly and varies based on each person's background, situation, and previous payment negotiations. These models suggest that demographics do impact a person's annual income. The data analysis performed was vital in determining which variables to include in the models. The models were produced based on the Naïve Bayes and Random Forest Methods, but there are many other machine learning methods that could be used. Trying different machine learning methods could yield a more accurate model. Additionally, the demographic information in the dataset was used to predict people's incomes, but these models could be made more accurate if additional data points were collected for each person.