STAT167 Final Written Report

JAMRR: Jiyeon Seo, Russel Wasko, Russell Ng, Arlette Jaime, Michelle Cheuk

Contents

What is your research question? Define the overall objective of your research project; then break down to 5-10 coherent research aims and/or sub-questions.

Overall Objective: Analysis on the correlation between the different demographic, personal backgrounds, and professional factors on the salary class of individuals.

Some coherent research aims and sub-questions we will answer can be seen below.

1. Is there a correlation between education and work class?
2. Does marital status have an effect on the amount of work time per week and salary
3. Does age and salary have a correlation?
4. Does gender affect the income of different types of jobs?
5. What jobs have the highest proportion of salaries > $50,000
6. Does race have an affect on an individual's income?
7. Does the amount of education an individual receives affect the salary class of that individual?
8. Can the proportions of race in education levels, explain why some races earn more than others?

Our dataset: Salary Prediction Classification

We decided on a dataset found from kaggle titled Salary Prediction Classification (link below). This dataset has 15 columns and has 35,561 observations. The columns are listed below with their explanations. We believe the data is relatively clean and has enough information to answer all of our research questions.

Explanation of the variables of the dataset

1. age : continuous.
2. workclass: a general term to represent the employment status of an individual - Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
3. fnlwgt: this is the number of people that census believes the entry represents - continuous.
4. education: Preschool , 1st-4th , 5th-6th , 7th-8th , 9th , 10th , 11th , 12th , HS-grad , Prof-school , Assoc-acdm , Assoc-voc , Some-college , Bachelors , Masters , Doctorate
5. education-num: a number that describe your education status from preschool to doctorate.
6. marital-status: marital status of an individual. Married-civ-spouse corresponds to a civilian spouse while Married-AF-spouse is a spouse in the Armed Forces. -Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.
7. occupation: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.
8. relationship: represents what this individual is relative to other Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.
9. race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
10. sex: Female, Male.
11. capital-gain: continuous.

12. capital-loss: continuous.
13. hours-per-week: continuous.
14. native-country: United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands.
15. salary: <=50K or >50K

```
# Import Data
dat = read.csv("salary.csv")
# Check the structure of the data
str(dat)
```

Data Cleaning and Preprocessing

1. Removing redundant variables

Based on the dataset summary, most of the attributes are easy to understand, except for fnlwgt, which may be short for "final weight." However, without knowledge of how it was calculated or its intended meaning, it may be challenging to use in our analysis. Additionally, education.num already provides a person's education history, so the variable education is unnecessary. Similarly, marital.status indicates a person's family status, making the variable relationship redundant. Therefore, we will remove the fnlwgt, education, and relationship variables from the dataset.

```
# 1. Remove redundant variables: fnlwgt and relationship
dat = dat[,-c(3,8)]
```

```
head(dat)
```

```
##   age         workclass  education education.num      marital.status
## 1  39         State-gov  Bachelors            13       Never-married
## 2  50  Self-emp-not-inc  Bachelors            13  Married-civ-spouse
## 3  38           Private    HS-grad             9            Divorced
## 4  53           Private       11th             7  Married-civ-spouse
## 5  28           Private  Bachelors            13  Married-civ-spouse
## 6  37           Private    Masters            14  Married-civ-spouse
##           occupation   race     sex capital.gain capital.loss hours.per.week
## 1       Adm-clerical  White    Male         2174            0             40
## 2    Exec-managerial  White    Male            0            0             13
## 3  Handlers-cleaners  White    Male            0            0             40
## 4  Handlers-cleaners  Black    Male            0            0             40
## 5     Prof-specialty  Black  Female            0            0             40
## 6    Exec-managerial  White  Female            0            0             40
##   native.country salary
## 1  United-States  <=50K
## 2  United-States  <=50K
## 3  United-States  <=50K
## 4  United-States  <=50K
## 5           Cuba  <=50K
## 6  United-States  <=50K
```

---

2. Cleaning Outliers

We will use the Interquartile Rule to find outliers in the variables age and education.num. Then we will use box plot to find outliers in capital gain.

```
# 2. Clean the Outliers by using IQR method
## Cleaning outliers in age ##
# summary(dat$age)
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   17.00   28.00   37.00   38.58   48.00   90.00
Q1_age = 28
Q3_age = 48
IQR_age = Q3_age - Q1_age
#IQR = Q3 - Q1
IQR_age
```

```
## [1] 20
```

```
# Find lowest value (LowerWhisker = Q1 - 1.5 * IQR_age)
LowerW_age = Q1_age - (1.5*IQR_age)
LowerW_age
```

```
## [1] -2
```

```
# Find upper value (UpperWhisker = Q3 + 1.5 * IQR_age)
UpperW_age = Q3_age + 1.5 * IQR_age
UpperW_age
```

```
## [1] 78
```

```
# Find observations above 78 (as UpperW_age =78)
dat = subset(dat, age <= 78)

## Cleaning outliers in education.num ##
# summary(dat$education.num)
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    9.00   10.00   10.08   12.00   16.00
Q1_education.num  = 9
Q3_education.num  = 12
IQR_education.num = Q3_education.num  - Q1_education.num
IQR_education.num
```

```
## [1] 3
```

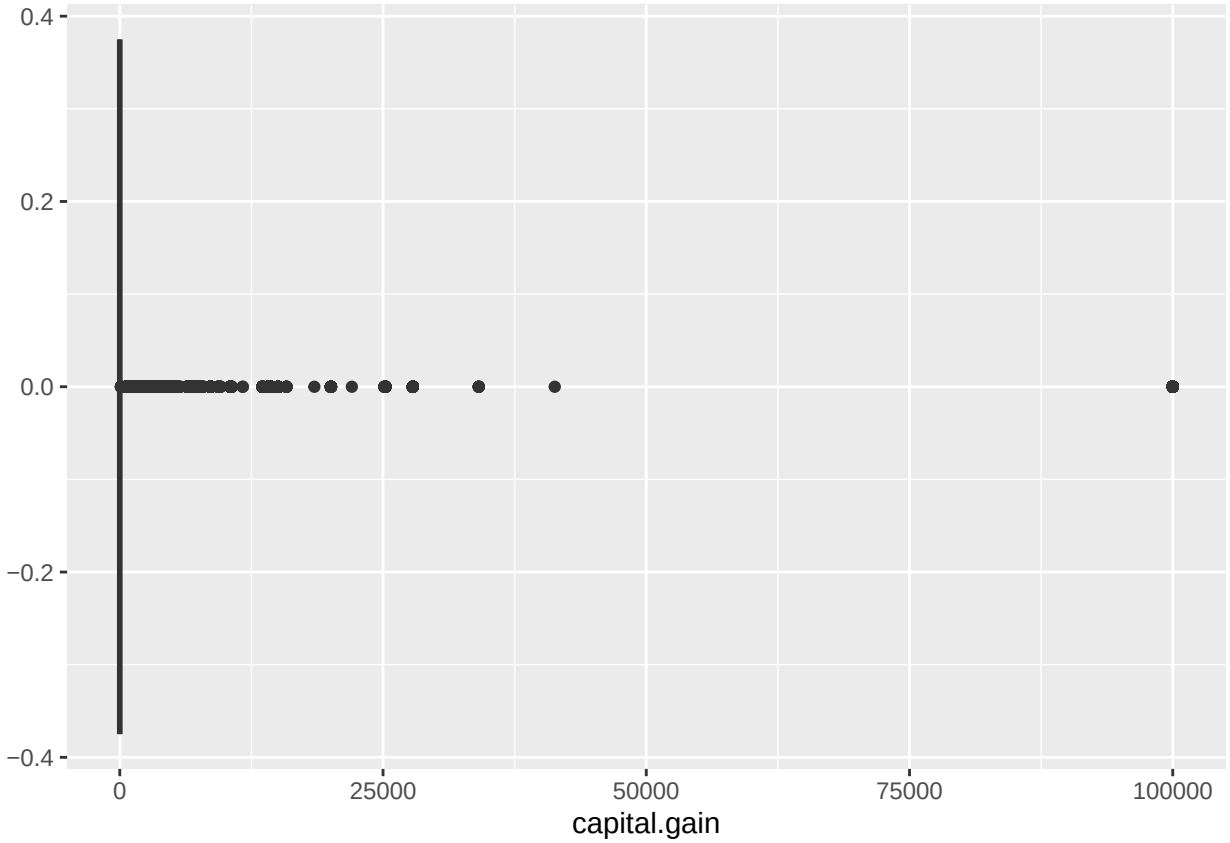
```
# Find lowest value (LowerWhisker = Q1 - 1.5 * IQR_education.num)
LowerW_education.num  = Q1_education.num - 1.5*IQR_education.num
LowerW_education.num
```

```
## [1] 4.5
```

```
# Find upper value: (UpperWhisker = Q3 + 1.5 * IQR_education.num)
UpperW_education.num  = Q3_education.num  + 1.5*IQR_education.num
UpperW_education.num
```

```
## [1] 16.5
```

```
# Find observations below 4.5
dat = subset(dat, education.num >= 4.5)

## Cleaning outliers in capital.gain ##
library(ggplot2)
# summary(dat$capital.gain)

box_plot = ggplot(dat, aes(x=capital.gain))+ geom_boxplot()
box_plot
```



99999 seems like a potential outlier, so we will remove it.

```
dat = subset(dat, capital.gain < 99999)
```

---

3. Reclassifying Categorical Variables and Min-Max normalization

If a categorical variable consists of too many field values, we should reclassify the field values of the categorical variables. For simplicity of the model, I will conduct the reclassification on four variables workclass, marital.status, native.country, and occupation.

```
# 3. Reclassifying Categorical Variables
## Change the "?" to Unknown ##
dat$occupation = gsub("?", "Unknown", dat$occupation, fixed = T )
dat$occupation = as.factor(dat$occupation)

dat$workclass = gsub("?", "Unknown", dat$workclass, fixed = T )
dat$workclass = as.factor(dat$workclass)

## Reclassify field values ##
## For marital.status ##
unique(dat$marital.status)
```

```
## [1] " Never-married"         " Married-civ-spouse"    " Divorced"
## [4] " Married-spouse-absent" " Separated"             " Married-AF-spouse"
## [7] " Widowed"
```

```
dat$marital.status[dat$marital.status == " Married-civ-spouse"] <- "Married"
dat$marital.status[dat$marital.status == " Married-spouse-absent"] <- "Married"
dat$marital.status[dat$marital.status == " Married-AF-spouse"] <- "Married"
unique(dat$marital.status)
```

```
## [1] " Never-married" "Married"        " Divorced"      " Separated"
## [5] " Widowed"
```

```
## For workclass ##
# Grouping "Federal-gov" "Local-gov", and "State-gov" into "Gov"
levels(dat$workclass)
```

```
## [1] " Federal-gov"      " Local-gov"        " Never-worked"
## [4] " Private"          " Self-emp-inc"     " Self-emp-not-inc"
## [7] " State-gov"        " Unknown"          " Without-pay"
```

```
levels(dat$workclass)[c(1,2,7)] = 'Gov'
# levels(dat$workclass)
levels(dat$workclass)[4:5] = 'Self-emp'
# levels(dat$workclass)
```

```
##
##                          ?                   Cambodia
##                1.714819e-02                5.147674e-04
##                     Canada                      China
##                3.732064e-03                2.219934e-03
##                    Columbia                       Cuba
##                1.769513e-03                2.606010e-03
##          Dominican-Republic                    Ecuador
##                1.608648e-03                8.043240e-04
##                 El-Salvador                    England
##                2.187761e-03                2.799048e-03
##                      France                    Germany
##                9.330159e-04                4.343350e-03
##                      Greece                  Guatemala
##                8.364970e-04                1.190400e-03
##                       Haiti          Holand-Netherlands
##                1.254746e-03                3.217296e-05
##                    Honduras                       Hong
##                3.539026e-04                5.791133e-04
##                     Hungary                      India
##                3.860755e-04                3.088604e-03
##                        Iran                    Ireland
##                1.383437e-03                7.721511e-04
##                       Italy                    Jamaica
##                1.769513e-03                2.573837e-03
##                       Japan                       Laos
##                1.962551e-03                4.504215e-04
##                      Mexico                  Nicaragua
##                1.158227e-02                9.651889e-04
##  Outlying-US(Guam-USVI-etc)                       Peru
##                4.504215e-04                9.651889e-04
```

```
##              Philippines                  Poland
##             5.951998e-03             1.737340e-03
##                 Portugal              Puerto-Rico
##             7.721511e-04             2.959912e-03
##                 Scotland                    South
##             3.860755e-04             2.477318e-03
##                   Taiwan                 Thailand
##             1.608648e-03             5.791133e-04
##          Trinadad&Tobago            United-States
##             5.469404e-04             9.092401e-01
##                  Vietnam                Yugoslavia
##             1.994724e-03             4.825944e-04
```

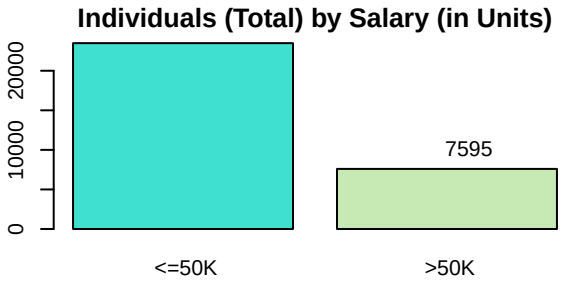
```r
# levels(dat$occupation)
```

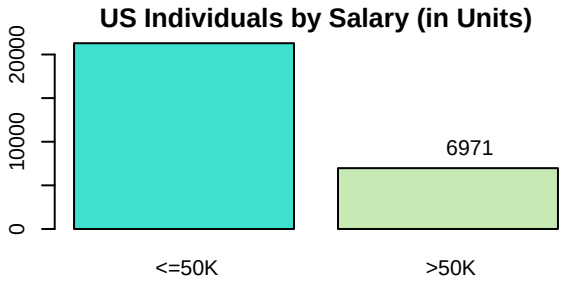
Since our data consists of both categorical and numeric variables,therefore, we will apply the min-max normalization to scale the numeric data. The cleaned dataset will be named as datnorm.

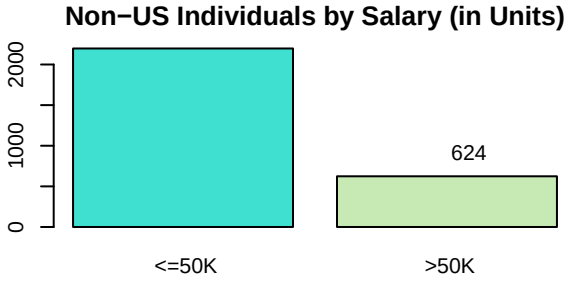
```r
# Min-Max normalization

datnorm <- dat
for (i in c(1, 4, 9, 10, 11)){
  mindf = min(datnorm[,i])
  maxdf = max(datnorm[,i])
  datnorm[,i] =(datnorm[,i] - mindf)/(maxdf - mindf)
}
```

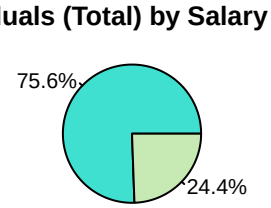
Exploratory Data Analysis

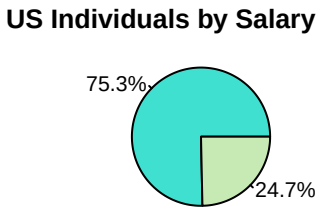
In this section, we will look more in depth into the data through visual graphs and data analysis, starting by analyzing the individual variables and move into how they interact with one another to help answer our sub questions.

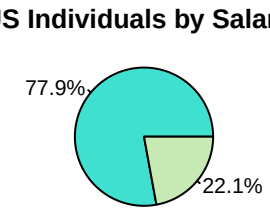










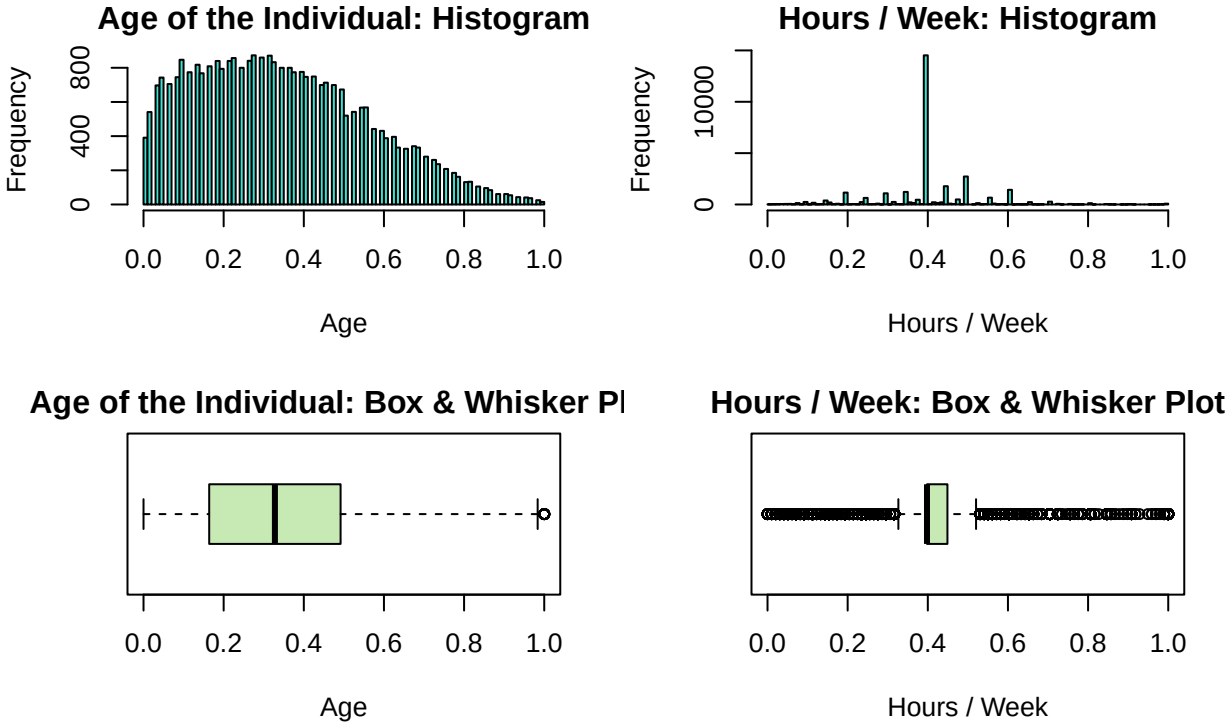


i) The dataset exhibits some class imbalance, with more than three-quarters of the records belonging to the <=50k salary segment. The segment representing salaries below 50k constitutes approximately 24% of the dataset.

ii) Additionally, it can be observed that the proportion of individuals earning over 50k salary is higher in the USA, accounting for around 24%, which aligns with the overall distribution.

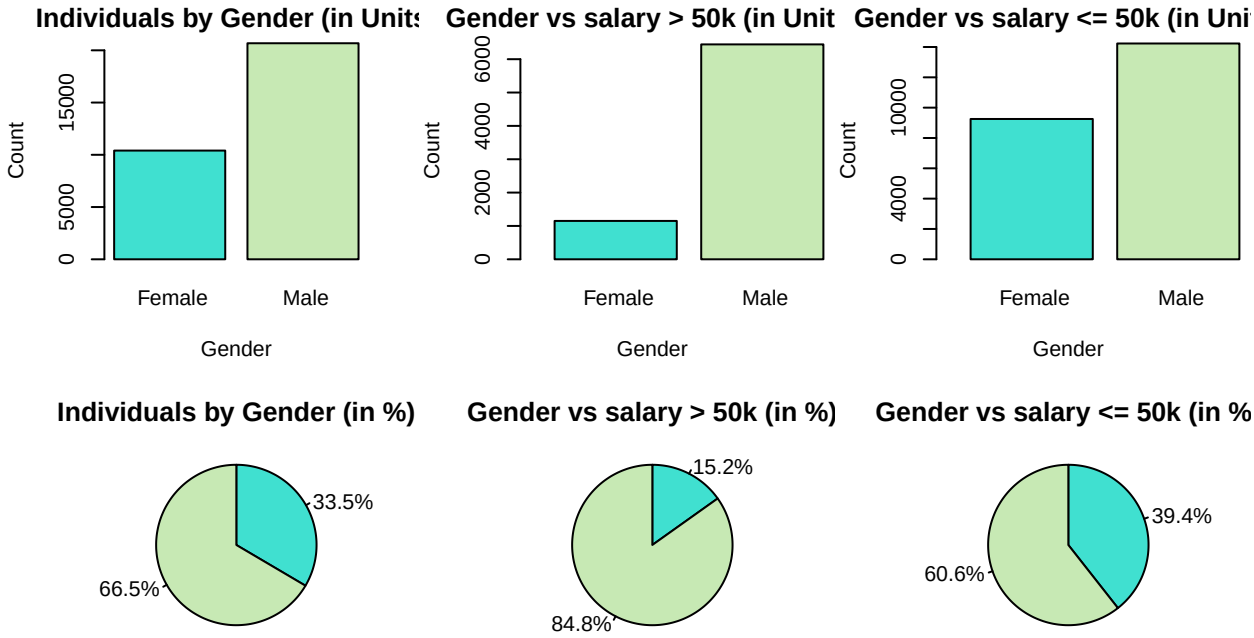
iii) In the NonUSA countries segment, the share of individuals earning below 50k salary is higher compared to the overall share (USA + NonUSA). It constitutes approximately 80% of the NonUSA segment.

Analysis of Numerical Variables



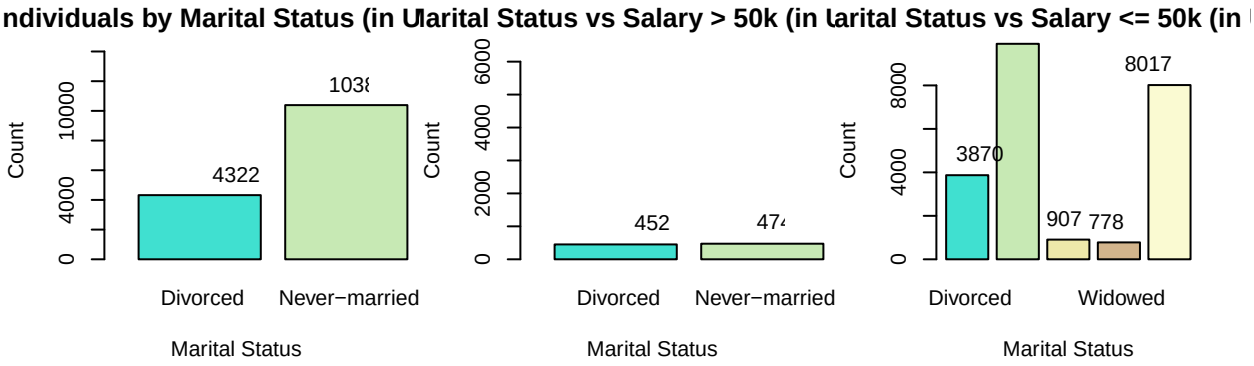
i) If we look at the age distribution, age around 20-50 work the most.
ii) People work around 40 hours/week the most.

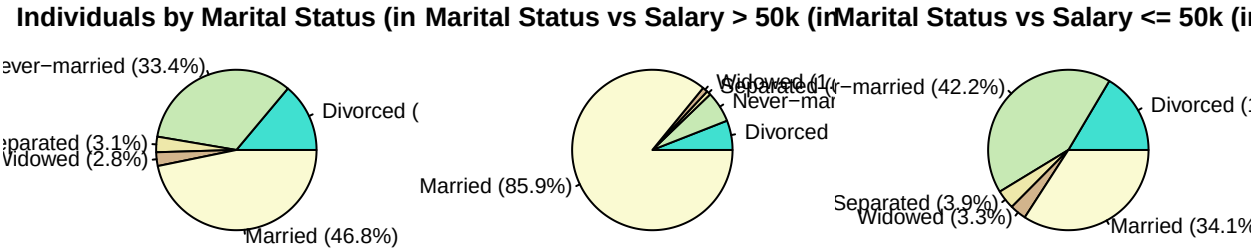
Analysis of Categorical Variables



There is a significant gender discrepancy in relation to salary.
i) Females represent only 15% of the individuals earning more than 50k, whereas they account for approximately 40% of those earning less than or equal to 50k. This suggests a gender disparity in salary distribution.
ii) Additionally, the gender disparity appears to be more pronounced in the USA, as the country comprises almost 90% of the total records in the dataset.

```
##
##       Divorced  Never-married      Separated        Widowed        Married
##            452            474             64             80           6525
```





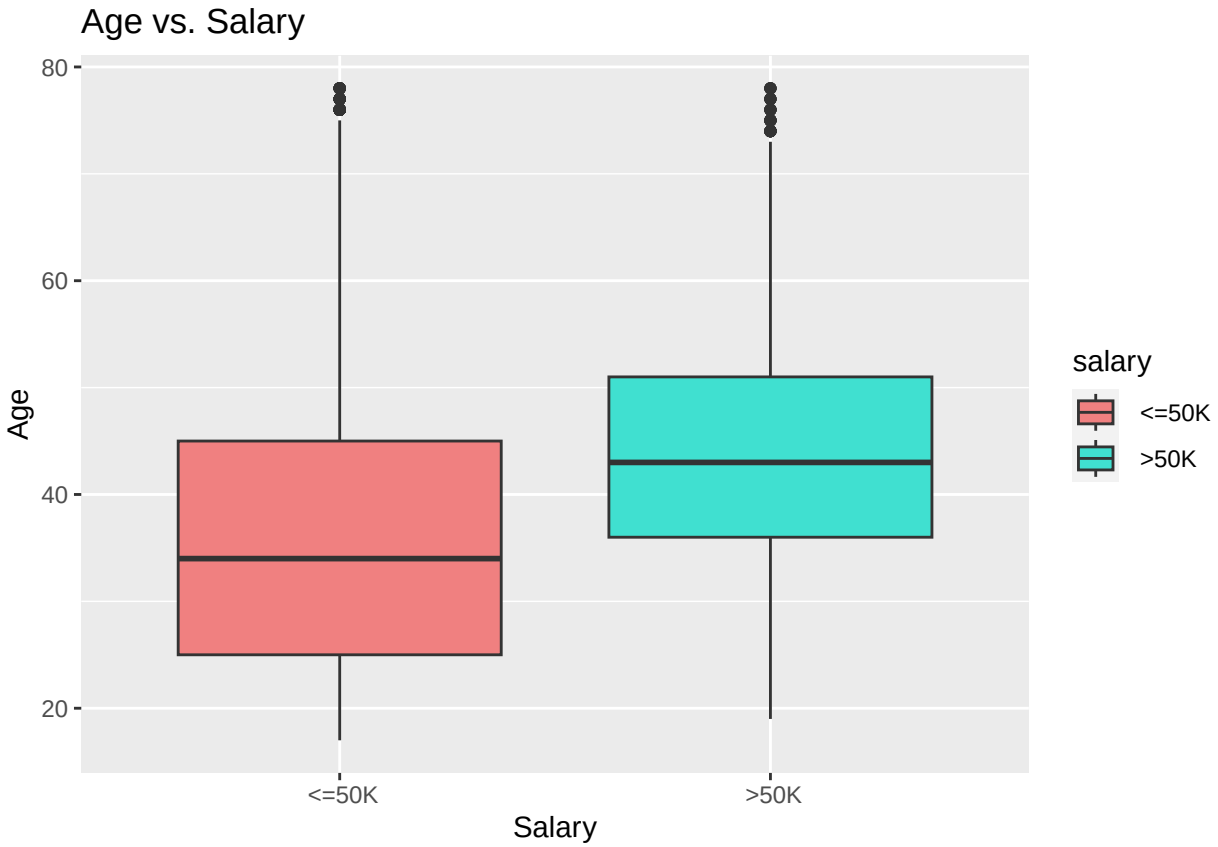
From the above plots, we can notice that Married people have higher salary, and comprise around 85% of total above 50K salaried individuals. And when we exclude the Married people from the overall population, most of them have a salary lower than 50K.

Sub-Questions

In this section, we will analyze and visualize various sub-questions pertaining to our original research question.

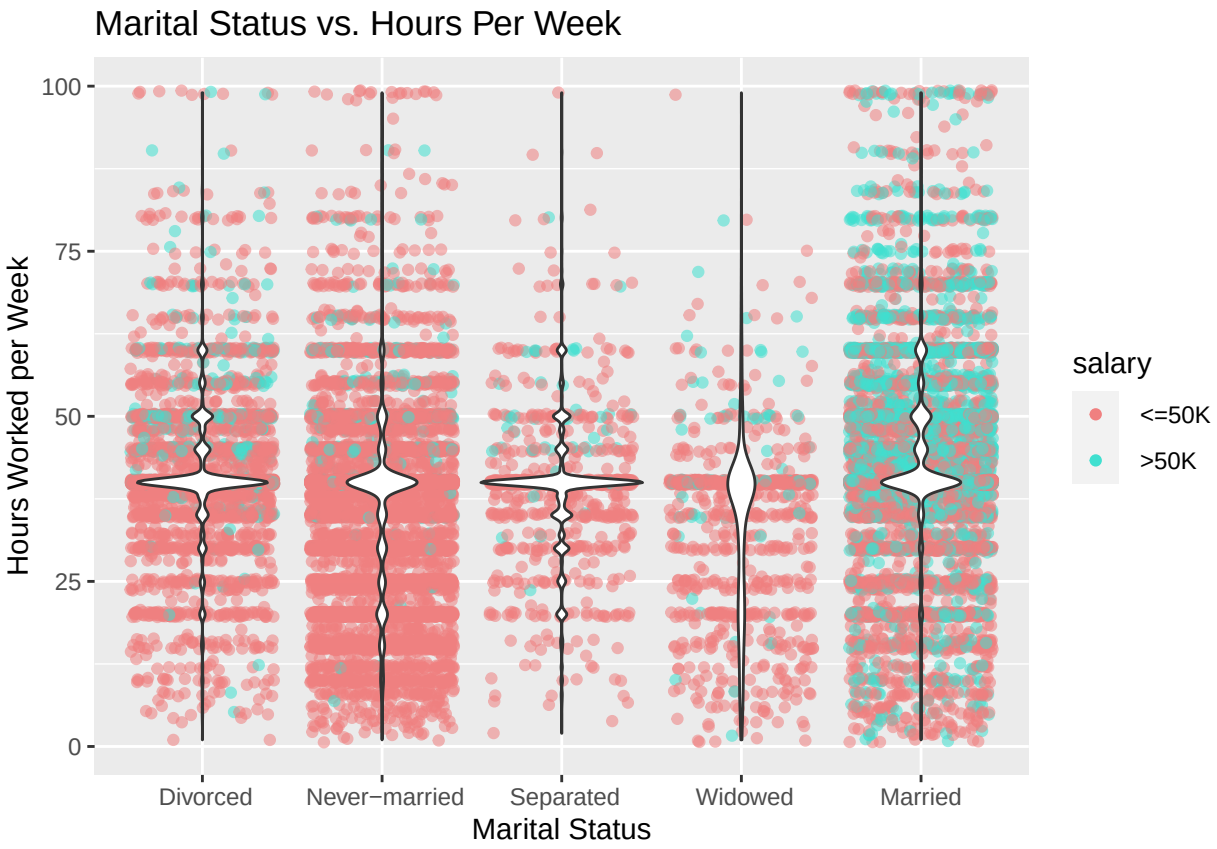
1. Is there a correlation between education and work class?

By overlaying a violin plot over a scatterplot of workclass vs. education num, we can see that in general, government workers, private workers, and self-employed workers all generally seek higher education.
9 = HS grad, 13 = Bachelors, 16 = Doctorate
i) There are very few government workers that do not have at least a high school diploma
ii) Private workers range in education from less than a high school diploma to higher education



2. Does marital status have an effect on the amount of work time per week and salary?

Boxplots of age grouped by the salary display that in general, people who are older are more likely to make more than 50,000 compared to their younger counterparts, with the average age of people making over 50,000 at 43 while those making under 50,000 are an average age of 34.



3. Does age and salary have a correlation?

Sorting people by marital status and overlaying a violin plot of age shows that most people regardless of marital status work 40 hours a week, which is the standard for most companies. There doesn't seem to be a large difference in working hours for those who are married, which is not what we expected, as we previously believed that those with families at home would work less hours.

We observed that married people work the same or more than their single counterparts while also making more money on average.

4. Does gender affect the income of different types of jobs?

```
job_gender_proportions <- dat %>%
  group_by(occupation, sex, salary) %>%
  summarise(count = n()) %>%
  group_by(occupation, sex) %>%
  mutate(proportion = count / sum(count)) %>%
  ungroup() %>%
  select(occupation, sex, proportion, salary) %>%
  spread(salary, proportion)
```

```
## `summarise()` has grouped output by 'occupation', 'sex'. You can override using
## the `.groups` argument.
```

```
job_gender_proportions
```

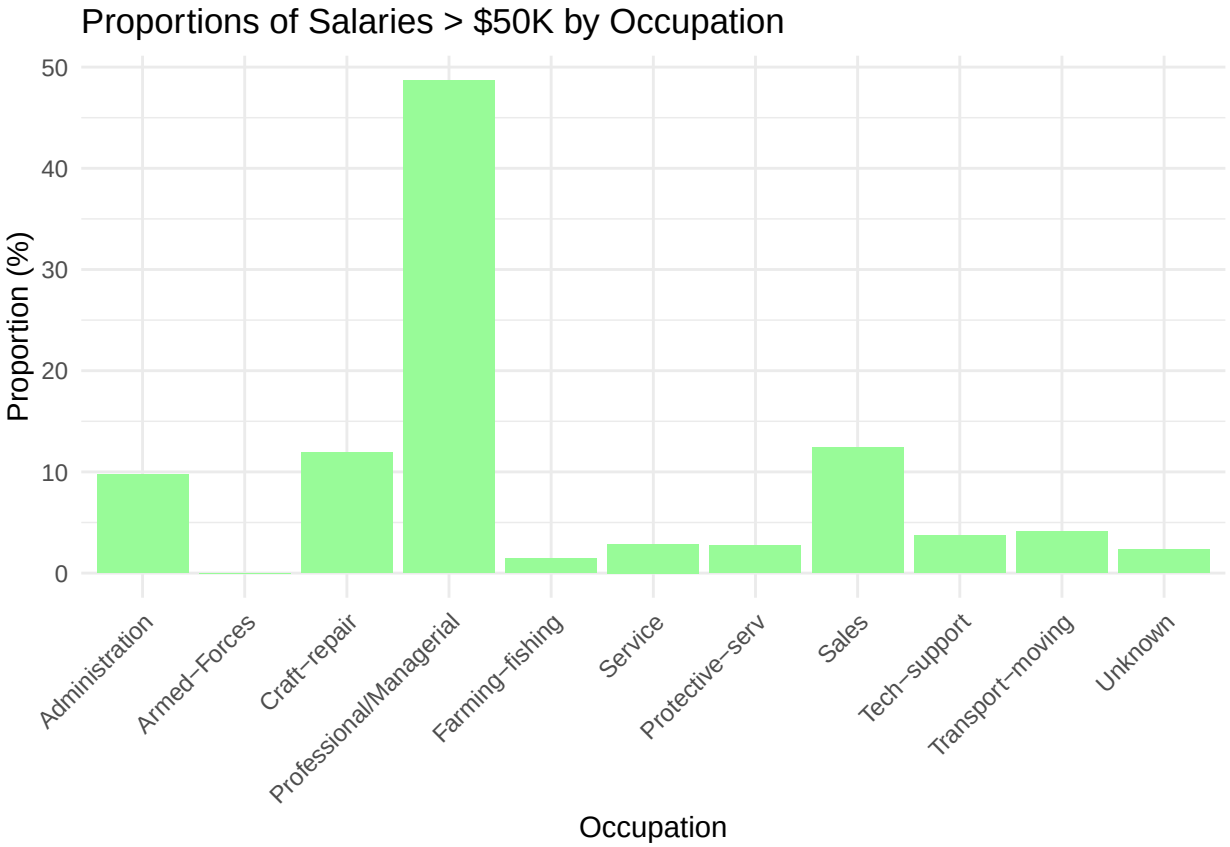
```
## # A tibble: 21 x 4
##    occupation                sex       ` <=50K` ` >50K`
##    <fct>                     <chr>        <dbl>   <dbl>
##  1 "Administration"          " Female"    0.925  0.0745
##  2 "Administration"          " Male"      0.798  0.202
##  3 " Armed-Forces"           " Male"      0.889  0.111
##  4 " Craft-repair"           " Female"    0.911  0.0892
##  5 " Craft-repair"           " Male"      0.759  0.241
##  6 "Professional/Managerial" " Female"    0.754  0.246
##  7 "Professional/Managerial" " Male"      0.433  0.567
##  8 " Farming-fishing"        " Female"    0.964  0.0357
##  9 " Farming-fishing"        " Male"      0.867  0.133
## 10 "Service"                 " Female"    0.972  0.0283
## # i 11 more rows
```

5. What jobs have the highest proportion of salaries > $50,000

```r
# 5

dat_50k <- dat[trimws(dat$salary) == ">50K", ]

dat_prop <- dat_50k %>%
  group_by(occupation) %>%
  summarise(proportion = n() / nrow(dat_50k) * 100) %>%
  arrange(desc(proportion))

head(dat_prop, 5)
```

```
## # A tibble: 5 x 2
##   occupation                proportion
##   <fct>                          <dbl>
## 1 "Professional/Managerial"      48.7
## 2 " Sales"                       12.4
## 3 " Craft-repair"                11.9
## 4 "Administration"                9.73
## 5 " Transport-moving"             4.08
```

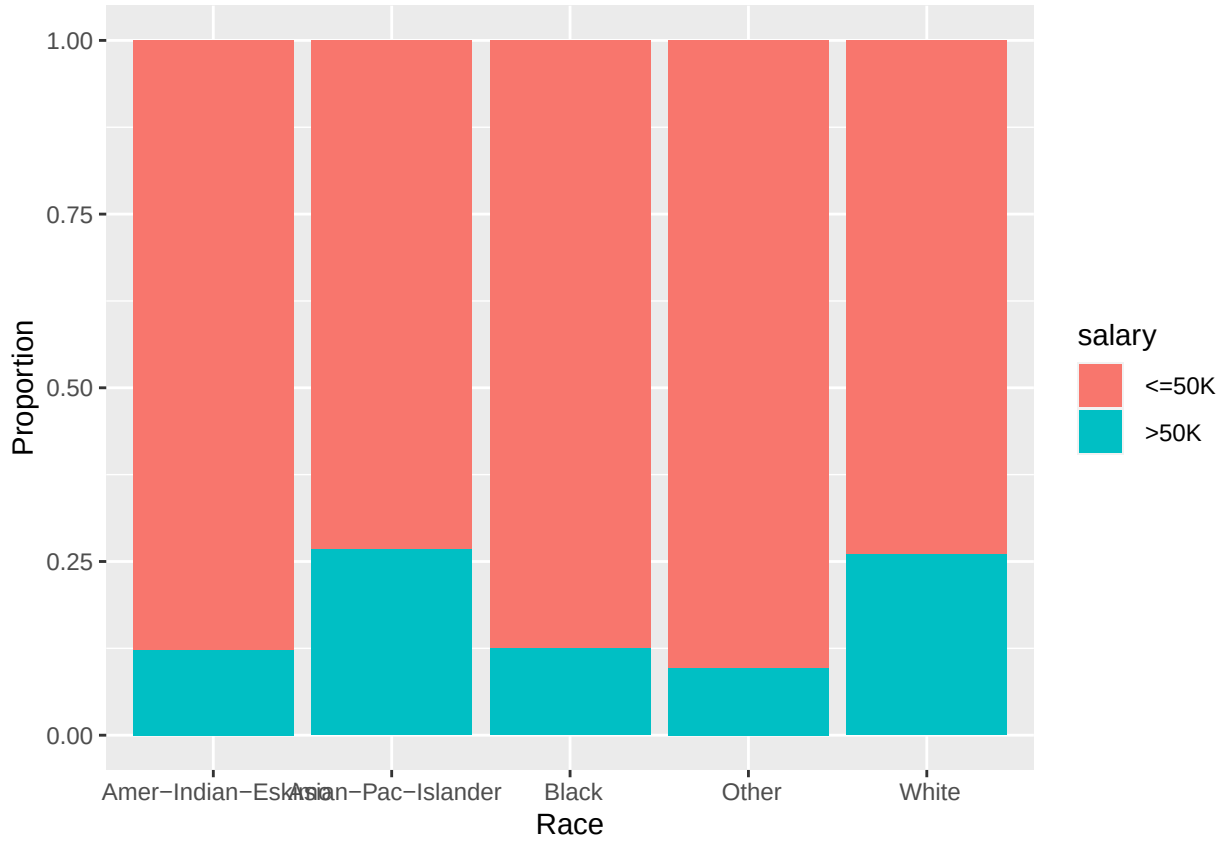
```r
ggplot(dat_prop, aes(x = occupation, y = proportion)) +
  geom_bar(stat = "identity", fill = "pale green") +
  labs(title = "Proportions of Salaries > $50K by Occupation",
       x = "Occupation", y = "Proportion (%)") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

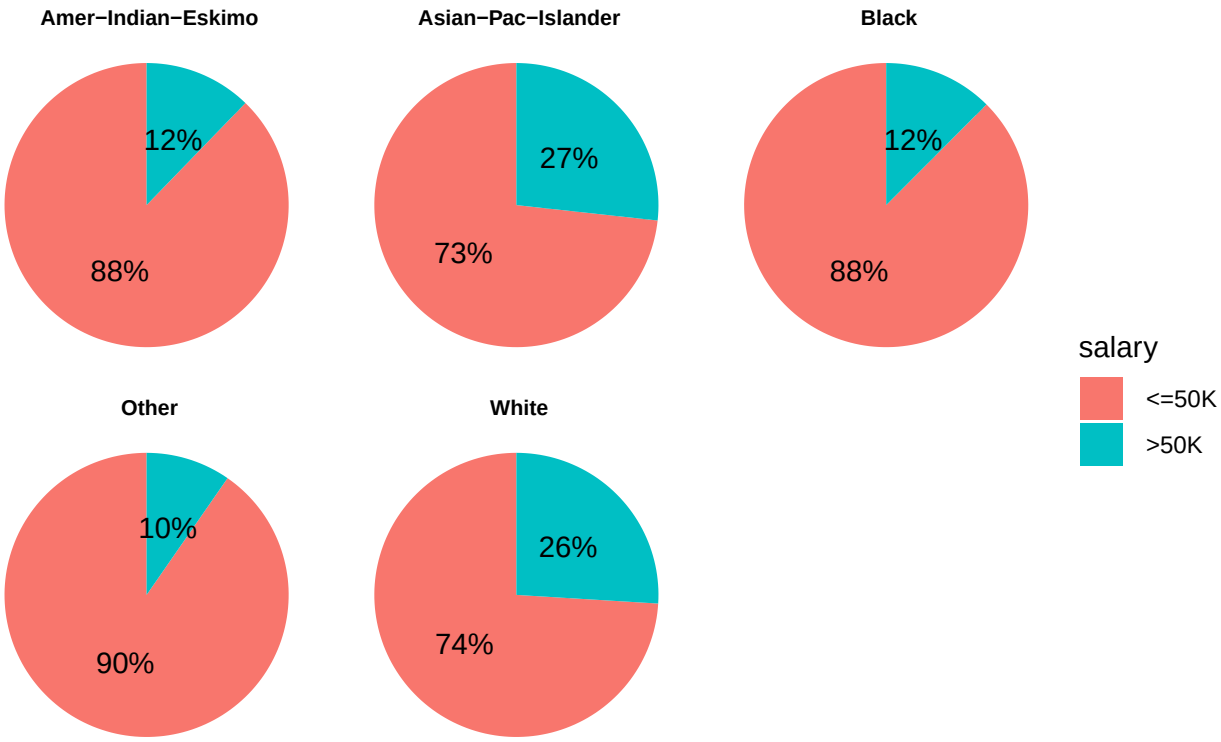


The jobs that have the highest proportion of salaries over $50,000 are professional/managerial, sales, craft repair, administration, and transport moving jobs.

Race vs. Salary

6. Does race have an affect on an individual's income?

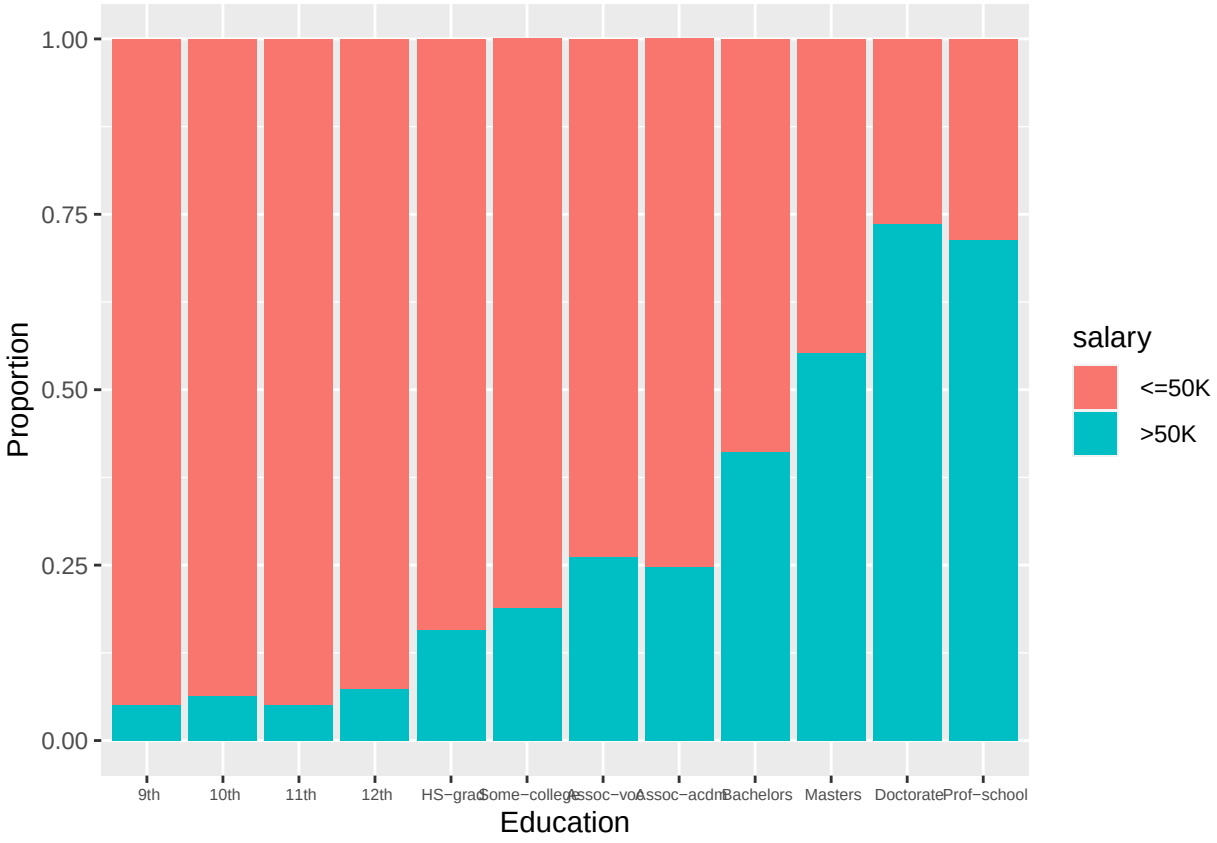


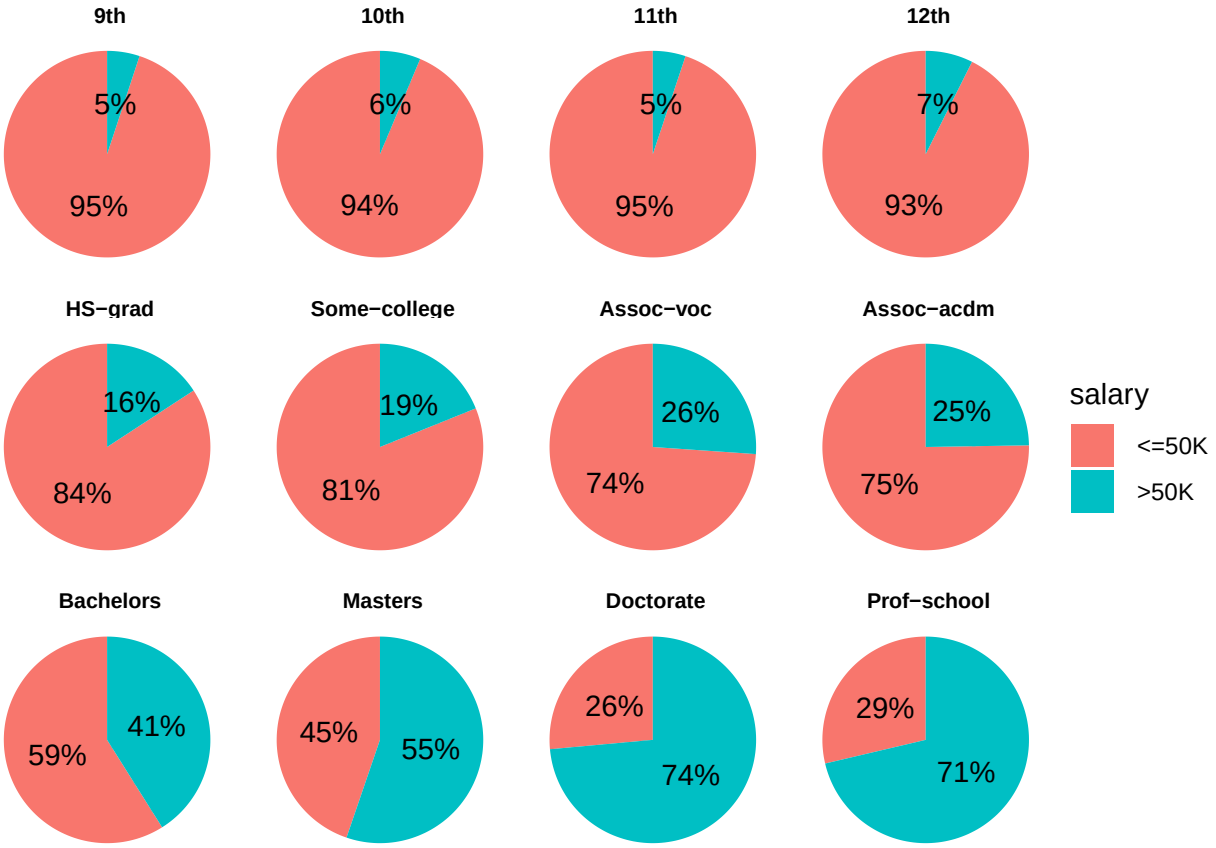


There are two races that have a considerably larger proportion that make over 50K. Asian-Pac-Islander people and White people have about 15% more people than all the other races.

Education vs. Salary

7. Does the amount of education an individual receives affect the salary class of that individual?





In order to answer this, I turned the education variable into an ordered factor starting with 9th grade level the lowest factor and going up from there. As the graphs show, individuals are more likely to earn more than 50K the higher education they have.

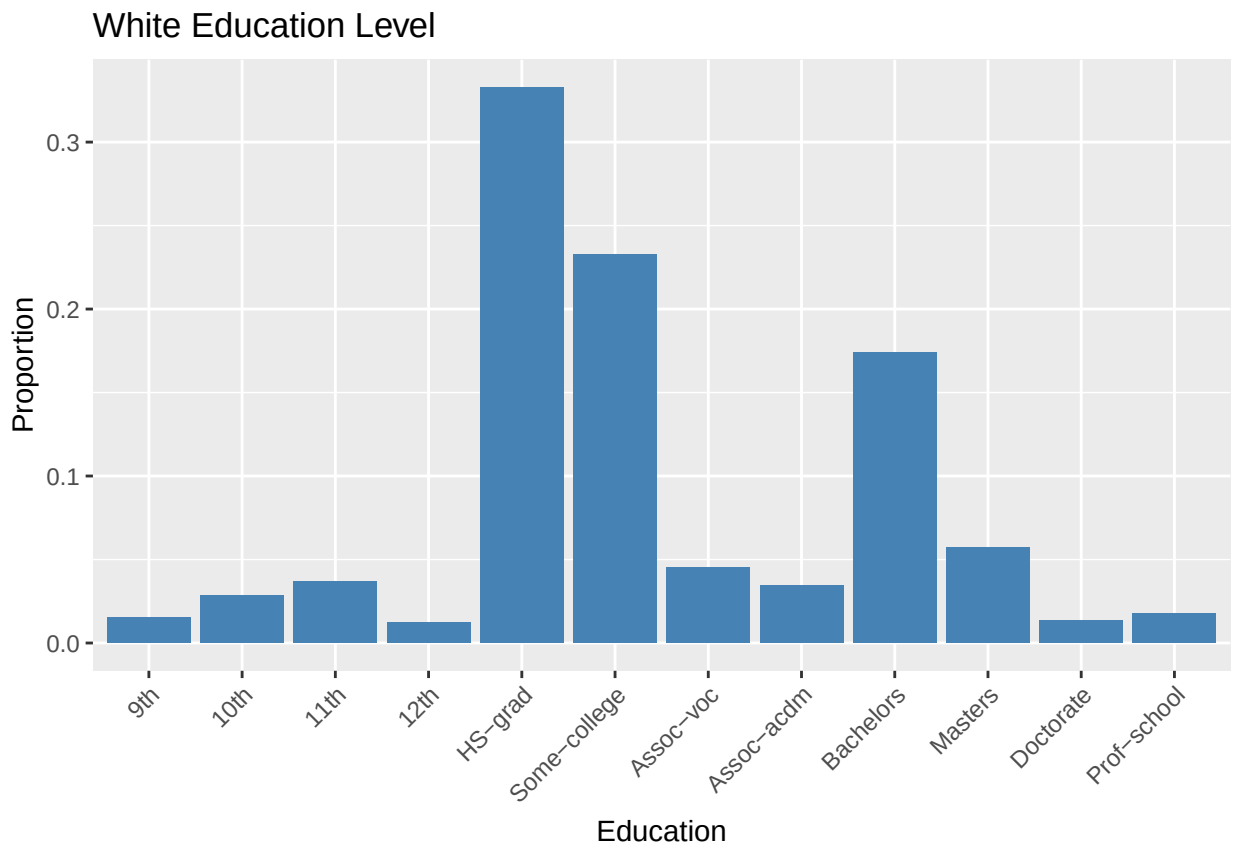
Race vs. Education

8. Can the proportions of race in education levels, explain why some races earn more than others?

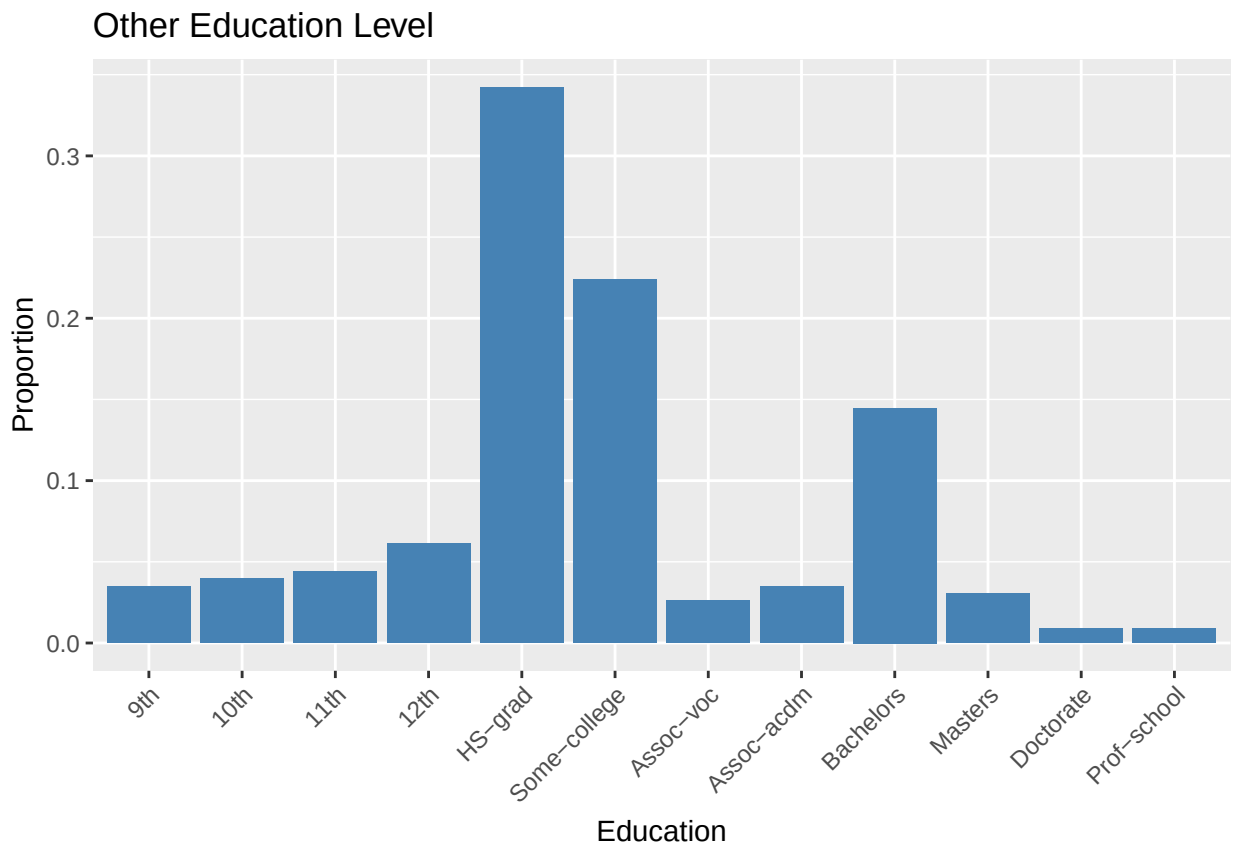
```
#Proportion of White people in each education level
educ_race_white <- dat %>%
  filter(race == " White") %>%
  group_by(education) %>%
  summarise(count = n()) %>%
  mutate(proportion = count / sum(count))

educ_race_white
```

```
## # A tibble: 12 x 3
##    education       count proportion
##    <ord>           <int>      <dbl>
##  1 " 9th"            401     0.0151
##  2 " 10th"           756     0.0285
##  3 " 11th"           973     0.0366
##  4 " 12th"           334     0.0126
##  5 " HS-grad"       8837     0.333
##  6 " Some-college"  6185     0.233
##  7 " Assoc-voc"     1201     0.0452
##  8 " Assoc-acdm"     913     0.0344
##  9 " Bachelors"     4629     0.174
## 10 " Masters"       1512     0.0569
## 11 " Doctorate"      353     0.0133
## 12 " Prof-school"    468     0.0176
```

```
#White education level barchart
ggplot(educ_race_white, aes(x = education, y = proportion)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  xlab("Education") +
  ylab("Proportion") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  ggtitle("White Education Level")
```




```
#Proportion of other races in each education level
educ_race_other <- dat %>%
   filter(race == " Other") %>%
  group_by(education) %>%
  summarise(count = n()) %>%
  mutate(proportion = count / sum(count))

#Other education level barchart
ggplot(educ_race_other, aes(x = education, y = proportion)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  xlab("Education") +
  ylab("Proportion") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  ggtitle("Other Education Level")
```



In order to see if the proportions of race in each education level might affect the race vs salary output. As the charts show, a bigger proportion of White people are able to go further in their education leading to more of them earning more than 50K compared to other races.

Partition the Dataset

Next, we make a copy of the dataset datnorm and name it as DF. Then we split the dataset DF into a training and a test set randomly by using sample() function. The training set DF.training contains 75% of the observations while the test data DF.test is the remaining 25% of observations.

```
# Creating training and test data set
# Divide the dataset into 2 portions in the ratio of 75: 25 for the training and test data set respecti
DF=datnorm
set.seed(123)
samp = sample(1:nrow(DF),round(0.75*nrow(DF)))
DF.training = DF[samp,]
DF.test= DF[-samp,]

# dim(DF.training)
# dim(DF.test)
```

To handle categorical variables in the KNN model, we utilize the dummyVars() function from the Caret library. This function helps us create indicator variables for the categorical variables such as workclass, marital.status, occupation, race, sex, native.country, and income. We apply this function separately to the training set and the test set. By doing so, the target variable income is also transformed into indicator variables. In this process, the income column is split into two columns: income less than 50K and income greater than 50K. We retain only the income greater than 50K column and rename it as income.more.50k. Consequently, the higher income class is represented by a value of 1, while the lower income class (below 50K annually) is represented by a value of 0. The resulting transformed dataset is referred to as SET B, with the training set and test set named training.dmy and test.dmy, respectively.

```
## DF.training - create dummy variable ##
dmy.training = dummyVars(" ~ .", data = DF.training)
training.dmy = data.frame(predict(dmy.training, newdata = DF.training))

# Dummy variables are created. We have to remove the income<=50k column.
training.dmy = training.dmy[-37] # Remove column 37 which is income <=50k
names(training.dmy)[names(training.dmy) == "income..50K"] = "income.more.50k" # Rename the income>50k c

## DF.test - create dummy variable ##
dmy.test = dummyVars(" ~ .", data = DF.test)
test.dmy = data.frame(predict(dmy.test, newdata = DF.test))

# Add the original income.class variable into the table
test.dmy = test.dmy[-37] # Remove column 37 which is income<=50
names(test.dmy)[names(test.dmy) == "income..50K"] = "income.more.50k" # Rename the income>50k column
```

Regression Model

Because our dataset deals with a Binary Dependent variable, we decided to work on developing a logistic regression model.

First we needed to mutate the salary character column to create a new salary binary variable where if " <=50k" = 0, and if " >50k" is set to 1.

Following this we started by running a base logistic regression model with every variable in our dataset and the output is shared below.

```
#Set salary to binary
datnorm <- datnorm %>%
  mutate(salary_binary = ifelse(salary == " <=50K", 0, 1))

#Base model with every variable
log_reg <- glm(salary_binary ~ age + workclass + education.num+ marital.status +occupation+ race+ sex+ c

summary(log_reg)
```

```
##
## Call:
## glm(formula = salary_binary ~ age + workclass + education.num +
##     marital.status + occupation + race + sex + capital.gain +
##     capital.loss + hours.per.week + native.country, family = binomial(),
##     data = datnorm)
##
## Coefficients: (1 not defined because of singularities)
##                                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)                       -7.474009   0.263775 -28.335  < 2e-16 ***
## age                                1.747686   0.100127  17.455  < 2e-16 ***
## workclass Never-worked            -9.721436 187.875443  -0.052  0.95873
## workclass Private                  0.063309   0.052217   1.212  0.22535
## workclassSelf-emp                 -0.205316   0.068461  -2.999  0.00271 **
## workclass Unknown                 -0.576562   0.118310  -4.873 1.10e-06 ***
## workclass Without-pay            -11.791292 128.800676  -0.092  0.92706
## education.num                      3.069672   0.104517  29.370  < 2e-16 ***
## marital.status Never-married      -0.500459   0.082368  -6.076 1.23e-09 ***
## marital.status Separated          -0.117199   0.159754  -0.734  0.46318
## marital.status Widowed             0.003872   0.152738   0.025  0.97977
## marital.statusMarried              2.109012   0.066144  31.885  < 2e-16 ***
## occupation Armed-Forces           -0.502596   1.525690  -0.329  0.74184
## occupation Craft-repair            0.087636   0.066029   1.327  0.18443
## occupationProfessional/Managerial  0.752347   0.059467  12.652  < 2e-16 ***
## occupation Farming-fishing        -1.125580   0.135713  -8.294  < 2e-16 ***
## occupationService                 -0.723401   0.090731  -7.973 1.55e-15 ***
## occupation Protective-serv         0.567465   0.116806   4.858 1.18e-06 ***
## occupation Sales                   0.333060   0.070097   4.751 2.02e-06 ***
## occupation Tech-support            0.736398   0.101616   7.247 4.27e-13 ***
## occupation Transport-moving       -0.096645   0.088909  -1.087  0.27703
## occupation Unknown                       NA         NA      NA       NA
## race Asian-Pac-Islander            0.608548   0.244695   2.487  0.01288 *
## race Black                         0.461659   0.230029   2.007  0.04475 *
## race Other                        -0.015916   0.354122  -0.045  0.96415
## race White                         0.622779   0.219875   2.832  0.00462 **
```

```
## sex Male                            0.212268   0.050543   4.200 2.67e-05 ***
## capital.gain                       13.224543   0.424409  31.160  < 2e-16 ***
## capital.loss                        2.918955   0.163082  17.899  < 2e-16 ***
## hours.per.week                      2.952316   0.157880  18.700  < 2e-16 ***
## native.countryUnited-States         0.311681   0.070645   4.412 1.02e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 34566  on 31081  degrees of freedom
## Residual deviance: 20760  on 31052  degrees of freedom
## AIC: 20820
##
## Number of Fisher Scoring iterations: 12
```

After creating the base model, we ran forward and backwards selection to ensure that our model is optimized with a variable selection process. It turns out that the forward and backwards selection kept the base model as is because every numeric variable was significant and every categorical variable contained at least one level that was significant in the model. The output is shared once again below.

```
#Forward and Backwards Selection
log_reg1 <- stepAIC(log_reg, direction = "both")
```

```
## Start:  AIC=20819.51
## salary_binary ~ age + workclass + education.num + marital.status +
##     occupation + race + sex + capital.gain + capital.loss + hours.per.week +
##     native.country
##
##                  Df Deviance   AIC
## <none>                20760 20820
## - race            4    20778 20830
## - sex             1    20777 20835
## - native.country  1    20779 20837
## - workclass       4    20790 20842
## - age             1    21068 21126
## - capital.loss    1    21090 21148
## - hours.per.week  1    21120 21178
## - occupation      9    21364 21406
## - education.num   1    21678 21736
## - capital.gain    1    22452 22510
## - marital.status  4    23990 24042
```

```
#Results
summary(log_reg1)
```

```
##
## Call:
## glm(formula = salary_binary ~ age + workclass + education.num +
##     marital.status + occupation + race + sex + capital.gain +
##     capital.loss + hours.per.week + native.country, family = binomial(),
##     data = datnorm)
##
```

```
## Coefficients: (1 not defined because of singularities)
##                                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)                       -7.474009   0.263775 -28.335  < 2e-16 ***
## age                                1.747686   0.100127  17.455  < 2e-16 ***
## workclass Never-worked            -9.721436 187.875443  -0.052  0.95873
## workclass Private                  0.063309   0.052217   1.212  0.22535
## workclassSelf-emp                 -0.205316   0.068461  -2.999  0.00271 **
## workclass Unknown                 -0.576562   0.118310  -4.873 1.10e-06 ***
## workclass Without-pay            -11.791292 128.800676  -0.092  0.92706
## education.num                      3.069672   0.104517  29.370  < 2e-16 ***
## marital.status Never-married      -0.500459   0.082368  -6.076 1.23e-09 ***
## marital.status Separated          -0.117199   0.159754  -0.734  0.46318
## marital.status Widowed             0.003872   0.152738   0.025  0.97977
## marital.statusMarried              2.109012   0.066144  31.885  < 2e-16 ***
## occupation Armed-Forces           -0.502596   1.525690  -0.329  0.74184
## occupation Craft-repair            0.087636   0.066029   1.327  0.18443
## occupationProfessional/Managerial  0.752347   0.059467  12.652  < 2e-16 ***
## occupation Farming-fishing        -1.125580   0.135713  -8.294  < 2e-16 ***
## occupationService                 -0.723401   0.090731  -7.973 1.55e-15 ***
## occupation Protective-serv         0.567465   0.116806   4.858 1.18e-06 ***
## occupation Sales                   0.333060   0.070097   4.751 2.02e-06 ***
## occupation Tech-support            0.736398   0.101616   7.247 4.27e-13 ***
## occupation Transport-moving       -0.096645   0.088909  -1.087  0.27703
## occupation Unknown                       NA         NA      NA       NA
## race Asian-Pac-Islander            0.608548   0.244695   2.487  0.01288 *
## race Black                         0.461659   0.230029   2.007  0.04475 *
## race Other                        -0.015916   0.354122  -0.045  0.96415
## race White                         0.622779   0.219875   2.832  0.00462 **
## sex Male                           0.212268   0.050543   4.200 2.67e-05 ***
## capital.gain                      13.224543   0.424409  31.160  < 2e-16 ***
## capital.loss                       2.918955   0.163082  17.899  < 2e-16 ***
## hours.per.week                     2.952316   0.157880  18.700  < 2e-16 ***
## native.countryUnited-States        0.311681   0.070645   4.412 1.02e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 34566  on 31081  degrees of freedom
## Residual deviance: 20760  on 31052  degrees of freedom
## AIC: 20820
##
## Number of Fisher Scoring iterations: 12
```

After developing our model we wanted analyze the accuracy of the model. First we developed the confusion matrix below. The model has a misclassification rate of 15.5%, a sensitivity (true positive rate) of .5842 and a specificity (true negative rate) of .9291.

```r
#Add predicted probabilities to the data frame
datnorm$predicted_prob <- predict(log_reg, type = "response")

#variable for predicted probabilities from model
probs <- predict(log_reg, type = "response")
```

```
#Predicted over or under .5
predicted <- ifelse(probs >= 0.5, 1, 0)


confusion_matrix <- table(datnorm$salary_binary,predicted)

#print confusion matrix
confusion_matrix
```

```
##    predicted
##         0     1
##   0 21820  1667
##   1  3158  4437
```

```
#misclassification rate
misclassRate <- mean(predicted != datnorm$salary_binary)

#print
misclassRate
```

```
## [1] 0.1552345
```

```
#True Positive
tp <- confusion_matrix[2,2]
#False Negative
fn <- confusion_matrix[2,1]
#True Negative
tn <- confusion_matrix[1,1]
#False Positive
fp <- confusion_matrix[1,2]

sensitivity <- tp/(tp+fn)
sensitivity
```

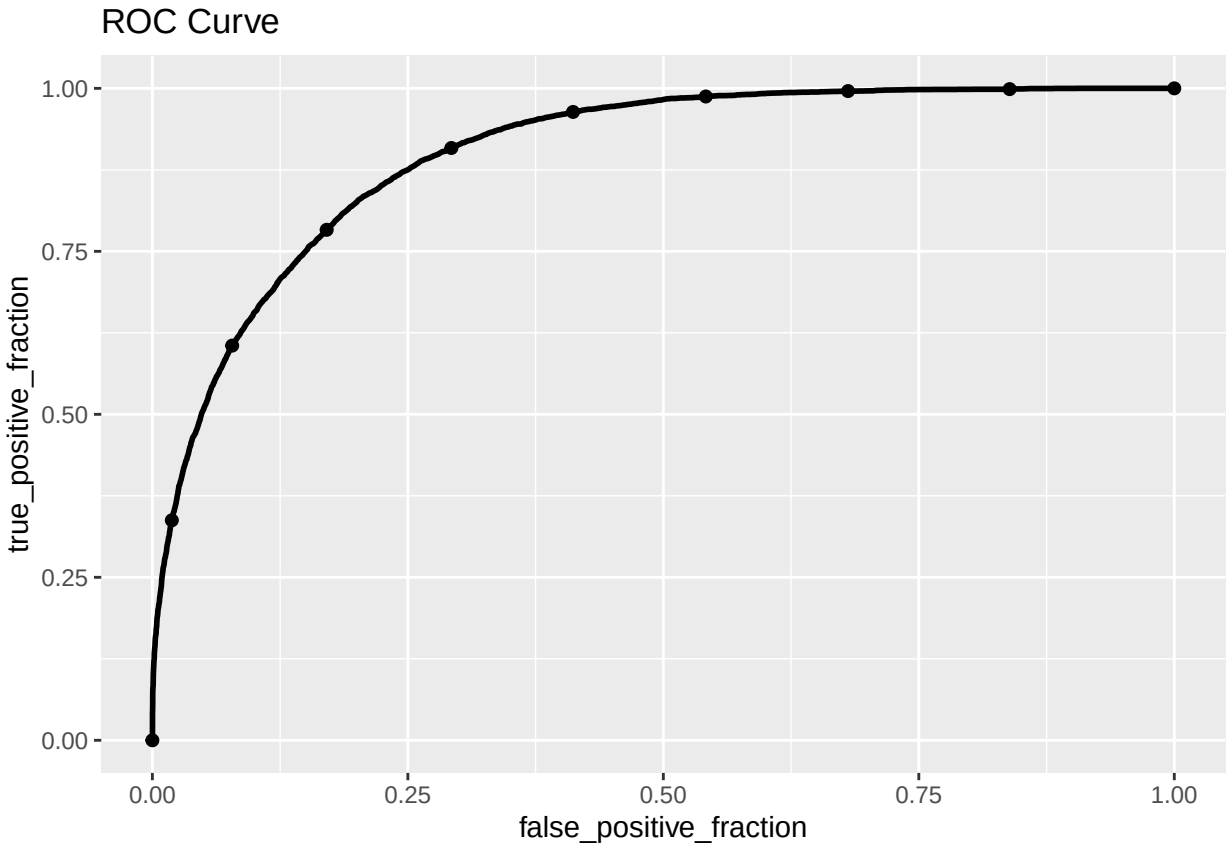
```
## [1] 0.5842001
```

```
specificity <- tn/(tn+fp)
specificity
```

```
## [1] 0.9290246
```

To visualize the sensitivity and specificity we plotted the ROC curve of our model below. The ROC curve leans towards the upper left portion of the plot and the AUC is .8986 which combine to show that our model is good at predicting results.

```
roc.df <- tibble(observed = datnorm$salary_binary,
                 predicted = probs)
#ROC
ggplot(data = roc.df, mapping = aes(d = observed, m = predicted)) +
  geom_roc(labels=F)+
  ggtitle("ROC Curve")
```




```
roc_data <- roc(roc.df$observed, roc.df$predicted)
```

```
## Setting levels: control = 0, case = 1
```

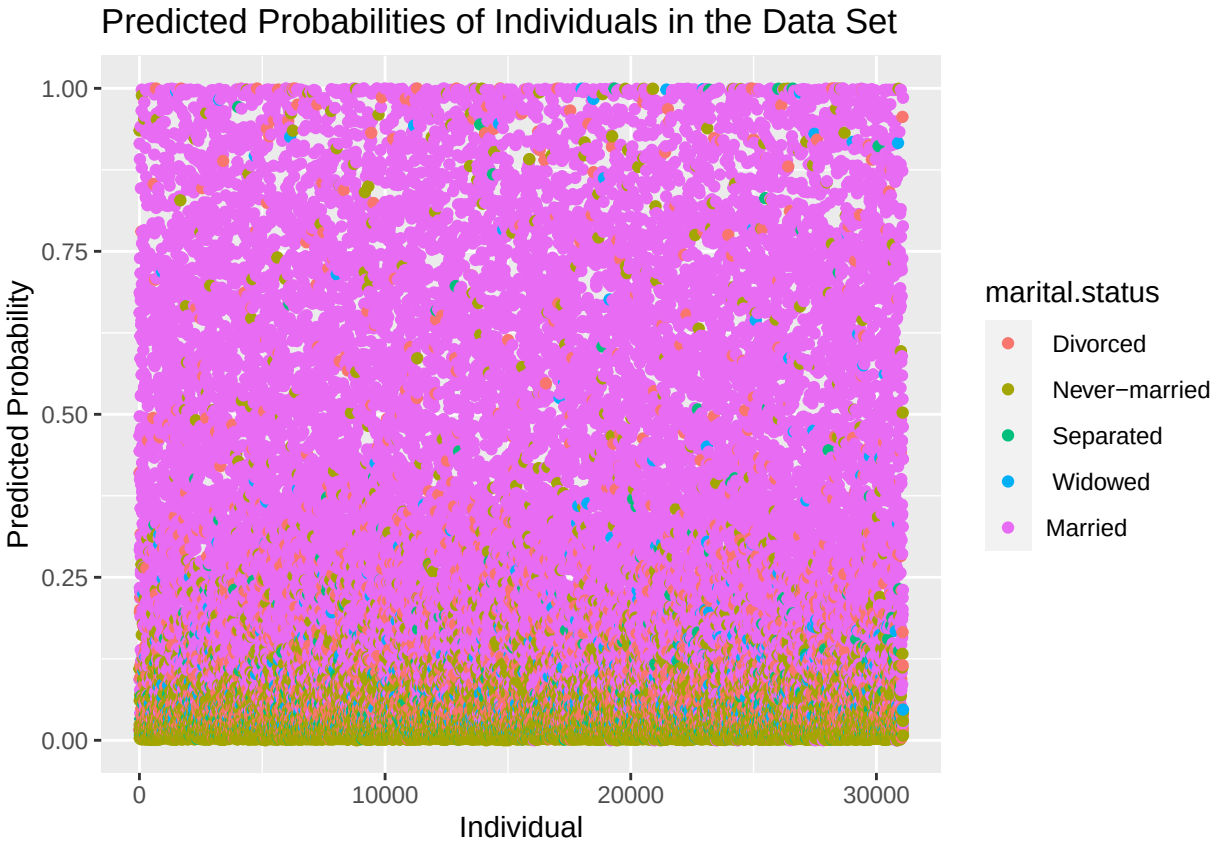
```
## Setting direction: controls < cases
```

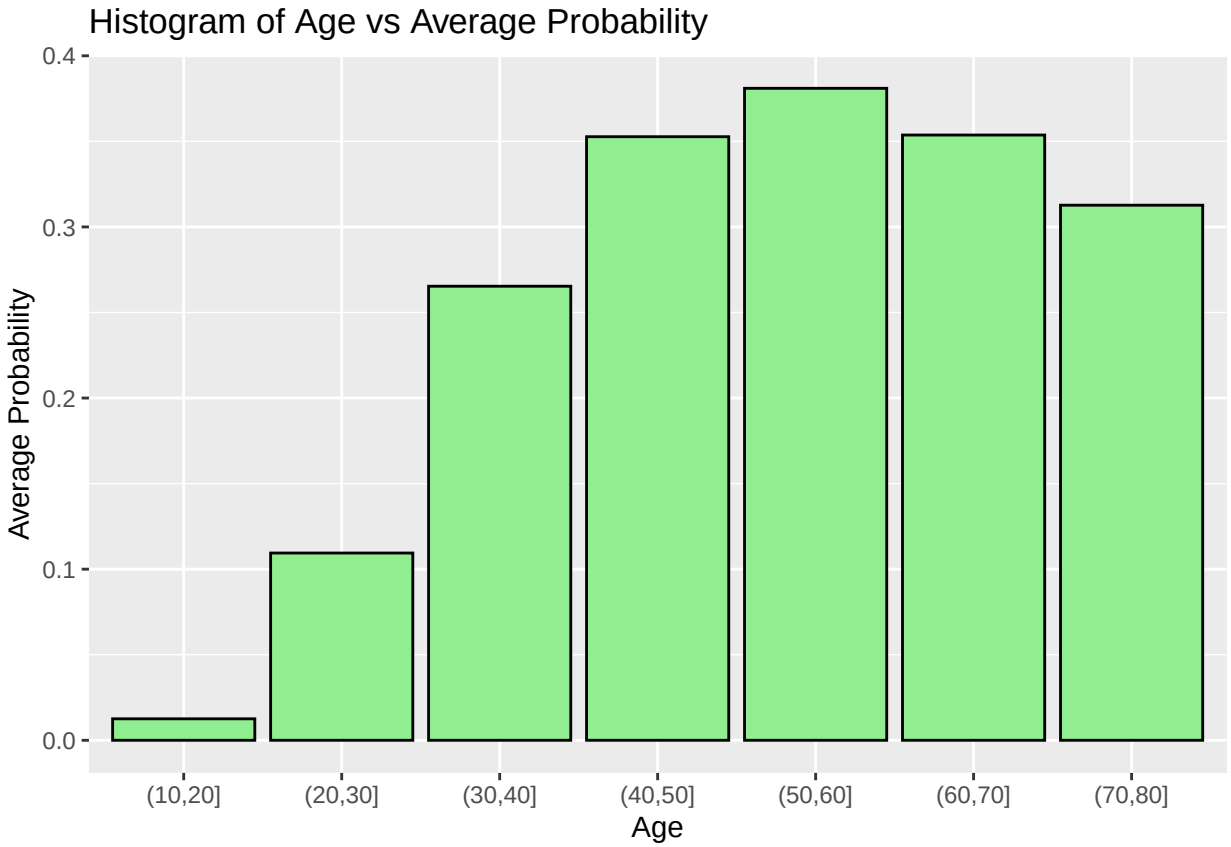
```
#Calculate AUC
auc <- auc(roc_data)

#Print AUC value
print(auc)
```

```
## Area under the curve: 0.8986
```

After validating the accuracy of the model we visualized some key findings from our results. The first plot displays the the predicted probability of each individual in our dataset to make a salary over 50,000 dollars. We colored the points by marital status as being married is the most significant categorical variable in our model. This significance can be seen as the never married individuals data points have the lowest probability of making 50,000 dollars while the married data points colored in purple have a significantly higher average predicted probability of making a salary of 50,000 dollars.

The next plot focuses on one of the most significant numeric variables in our model, age. The plot shows the average predicted probability of making 50,000 dollars for bins of every 10 years. It is evident that there is a trend that older individual had better predicted probabilities than younger individuals, with the peak age being 50-60 years old.


Predicted Probabilities of Individuals in the Data Set
Predicted Probability
1.00
0.75
0.50
0.25
0.00
0
10000
20000
30000
Individual
marital.status
Divorced
Never−married
Separated
Widowed
Married


Histogram of Age vs Average Probability
Average Probability
0.4
0.3
0.2
0.1
0.0
(10,20]
(20,30]
(30,40]
(40,50]
(50,60]
(60,70]
(70,80]
Age

KNN Analysis

```
## Confusion Matrix and Statistics
##
##          DF_test_labels
## estknn.20  <=50K  >50K
##     <=50K   5864    20
##     >50K       4  1882
##
##                Accuracy : 0.9969
##                  95% CI : (0.9954, 0.998)
##     No Information Rate : 0.7552
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.9916
##
##  Mcnemar's Test P-Value : 0.0022
##
##             Sensitivity : 0.9993
##             Specificity : 0.9895
##          Pos Pred Value : 0.9966
##          Neg Pred Value : 0.9979
##              Prevalence : 0.7552
##          Detection Rate : 0.7547
##    Detection Prevalence : 0.7573
##       Balanced Accuracy : 0.9944
##
##        'Positive' Class :  <=50K
##

## Confusion Matrix and Statistics
##
##          DF_test_labels
## estknn.10  <=50K  >50K
##     <=50K   5865    19
##     >50K       3  1883
##
##                Accuracy : 0.9972
##                  95% CI : (0.9957, 0.9982)
##     No Information Rate : 0.7552
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.9923
##
##  Mcnemar's Test P-Value : 0.001384
##
##             Sensitivity : 0.9995
##             Specificity : 0.9900
##          Pos Pred Value : 0.9968
##          Neg Pred Value : 0.9984
##              Prevalence : 0.7552
##          Detection Rate : 0.7548
##    Detection Prevalence : 0.7573
##       Balanced Accuracy : 0.9947
```

```
##
##        'Positive' Class :  <=50K
##


## Confusion Matrix and Statistics
##
##          DF_test_labels
## estknn.10  <=50K  >50K
##     <=50K   5865    19
##     >50K       3  1883
##
##                Accuracy : 0.9972
##                  95% CI : (0.9957, 0.9982)
##     No Information Rate : 0.7552
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.9923
##
##  Mcnemar's Test P-Value : 0.001384
##
##             Sensitivity : 0.9995
##             Specificity : 0.9900
##          Pos Pred Value : 0.9968
##          Neg Pred Value : 0.9984
##              Prevalence : 0.7552
##          Detection Rate : 0.7548
##    Detection Prevalence : 0.7573
##       Balanced Accuracy : 0.9947
##
##        'Positive' Class :  <=50K
##
```

With an accuracy rate of 83%, it is advisable to include most of the factors in the analysis. This is supported by the high sensitivity, which indicates a high proportion of correctly identified positive observations, and the high balanced accuracy. These factors contribute to the credibility of using the KNN model to determine the impact of other variables on salary.

Conclusion: By examining both the regression model and the KNN model, we have arrived at the conclusion that multiple factors significantly influence salary prediction. The accuracy of both models provides strong support for this conclusion, indicating that they have effectively captured the impact of various variables on salary. The exploratory data analysis also corroborates these conclusions that many factors affect the prediction of salary. Some of the factors within the model that were the most influential were:
- Marital Status
- Age
- Hours per Week

All of these variables were extremely influential to predict if someone will make over $50,000 per year, as shown through the logistic regression analysis and the exploratory data analysis.
Another finding that we observed was that Capital Gains and Losses were both positive to predict salary. This is not what we expected from the model, but it makes sense because capital gains and losses are the gains from the selling of assets. To be able to gain and lose money from assets, there needs to be capital in the first place to buy assets.
Additionally, we found that people that are not married were negative to the model, meaning marital status that are single, separated, or widowed earn less on average compared to their married counterparts.

Limitations:
Some limitations we faced regarding this data analysis was the fact that this data set was older, particularly from a 1994 Census. This means that this data would probably not be completely representative of what the current work landscape is like, especially since surveying was not as sophisticated as it is today.
Additionally, extraction of the data is likely to be skewed, especially the ratio of male to female and married to unmarried. Through our data analysis, we discovered that this dataset was mostly comprised of male workers, which may skew the data as it is not representative of the population.

Contributions

Russel Wasko: Helped build research questions, logistic regression
Russell Ng: Helped build research questions, exploratory data analysis, formatting of RMD file Michelle Cheuk: Helped build research questions, exploratory data analysis Arlette Jaime: Helped build research questions, exploratory data analysis
Jiyeon Seo: Helped build research questions, exploratory data analysis, KNN analysis and data cleaning

Link to the Dataset

Link to the Dataset
Google Drive Link