3 Analysis of Data Quality

```
library(ggplot2)
library(skimr)

#load data
H1B <- read.csv('H1B_26variable.csv')

#data dimension
dim(H1B)
```

```
## [1] 654360     27
```

Dataset Summary

```
#data summary
summary(H1B)
```

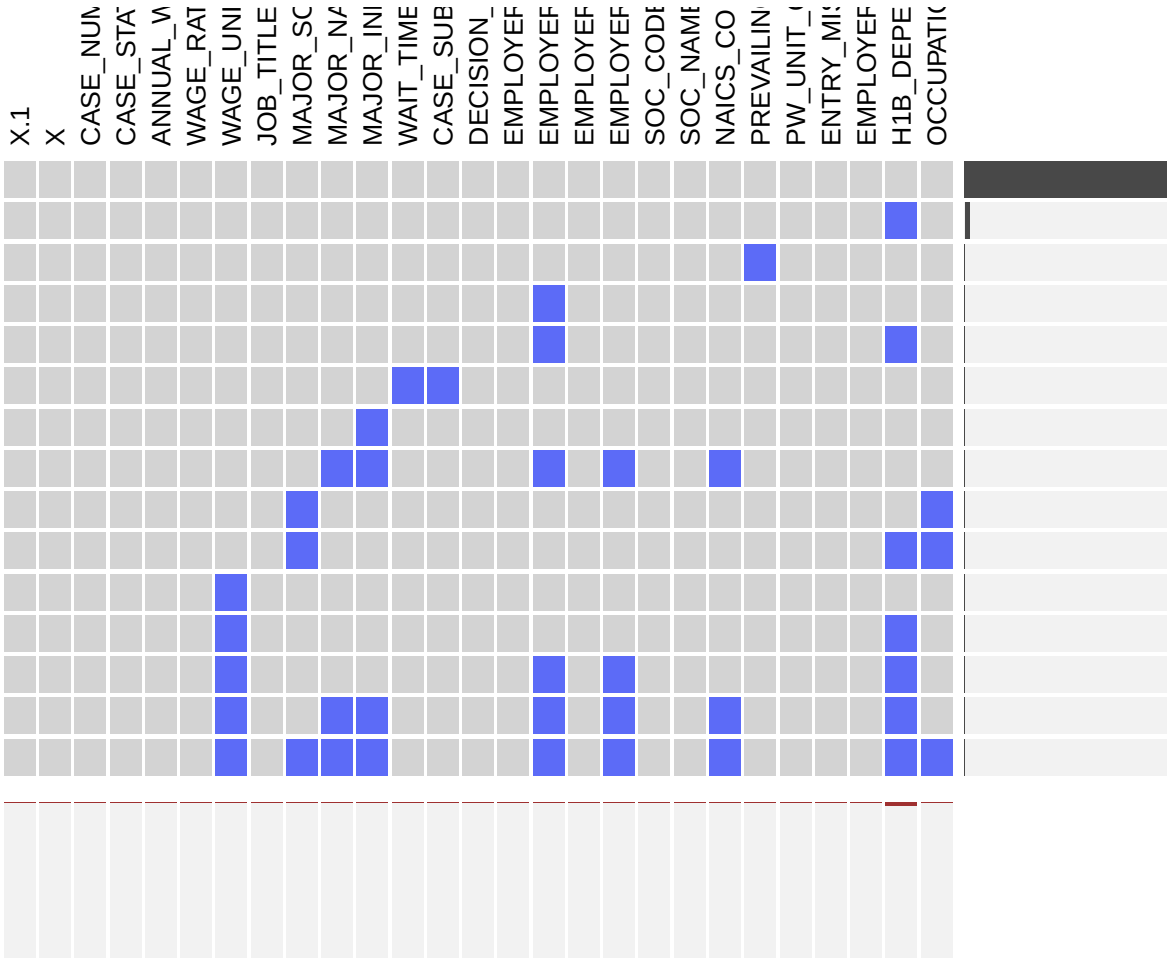
```
##       X.1                X                       CASE_NUMBER
##  Min.   :     1   Min.   :     1   I-200-09180-429413:     1
##  1st Qu.:163591   1st Qu.:163591   I-200-09308-748082:     1
##  Median :327180   Median :327180   I-200-09323-288635:     1
##  Mean   :327180   Mean   :327180   I-200-09351-355625:     1
##  3rd Qu.:490770   3rd Qu.:490770   I-200-09362-088372:     1
##  Max.   :654360   Max.   :654360   I-200-10060-038969:     1
##                                    (Other)           :654354
##               CASE_STATUS      ANNUAL_WAGE        WAGE_RATE_OF_PAY_FROM
##  CERTIFIED          :579449   Min.   :0.000e+00   Min.   :0.00e+00
##  CERTIFIED-WITHDRAWN: 45004   1st Qu.:7.191e+04   1st Qu.:7.00e+04
##  DENIED             :  8627   Median :8.670e+04   Median :8.56e+04
##  WITHDRAWN          : 21280   Mean   :9.632e+04   Mean   :9.12e+04
##                               3rd Qu.:1.094e+05   3rd Qu.:1.08e+05
##                               Max.   :1.000e+09   Max.   :1.00e+09
##
##   WAGE_UNIT_OF_PAY                          JOB_TITLE      MAJOR_SOC_CODE
##  Bi-Weekly:    99   SOFTWARE DEVELOPER         : 34907   Min.   :11.00
##  Hour     : 44371   SOFTWARE ENGINEER          : 31943   1st Qu.:15.00
##  Month    :   497   PROGRAMMER ANALYST         : 14109   Median :15.00
##  Week     :   150   SENIOR SOFTWARE ENGINEER   :  8430   Mean   :16.08
##  Year     :609230   SENIOR SYSTEMS ANALYST JC60:  7041   3rd Qu.:15.00
##  NA's     :    13   DEVELOPER                  :  6244   Max.   :53.00
##                     (Other)                    :551686   NA's   :596
##  MAJOR_NAICS_CODE
##  Min.   :10.00
##  1st Qu.:53.00
##  Median :54.00
##  Mean   :52.22
##  3rd Qu.:54.00
##  Max.   :99.00
##  NA's   :6
##                                          MAJOR_INDUSTRY
##  Professional, Scientific, and Technical Services:415563
```

```
##  Manufacturing                                    : 48614
##  Finance and Insurance                            : 38729
##  Information                                      : 37066
##  Educational Services                             : 34175
##  (Other)                                          : 80201
##  NA's                                             :    12
##    WAIT_TIME      CASE_SUBMITTED   DECISION_DATE
##  Min.   :   0.0   3/16/18: 11729   3/22/18: 16627
##  1st Qu.:   6.0   3/15/18: 11419   3/15/18: 14287
##  Median :   6.0   3/14/18: 11278   3/21/18: 11109
##  Mean   :  31.8   3/20/18: 10544   3/20/18: 11001
##  3rd Qu.:   6.0   3/13/18: 10354   3/26/18: 10249
##  Max.   :2794.0   (Other):599035   3/19/18: 10234
##  NA's   :1        NA's   :     1   (Other):580853
##          EMPLOYER_CITY    EMPLOYER_STATE   EMPLOYER_POSTAL_CODE
##  NEW YORK       : 32553   CA     :113307   19103  : 19643
##  CHICAGO        : 23206   NJ     : 83892   20850  : 15478
##  PHILADELPHIA   : 21319   TX     : 79902   75024  : 13740
##  PLANO          : 16757   NY     : 45783   77845  : 13095
##  ROCKVILLE      : 15712   IL     : 39231   7080   :  9392
##  COLLEGE STATION: 13466   (Other):292172   94043  :  9194
##  (Other)        :531347   NA's   :    73   (Other):573818
##                  EMPLOYER_COUNTRY     SOC_CODE
##  AUSTRALIA               :    29   15-1132:194777
##  BELGIUM                 :     3   15-1121: 74433
##  CANADA                  :    33   15-1199: 63973
##  INDIA                   :     1   15-1133: 27154
##  UNITED STATES OF AMERICA:654287   15-1131: 26886
##  NA's                    :     7   15-2031: 13380
##                                    (Other):253757
##                                SOC_NAME        NAICS_CODE
##  SOFTWARE DEVELOPERS, APPLICATIONS    :194439   Min.   :    23
##  COMPUTER OCCUPATIONS, ALL OTHER      : 60967   1st Qu.:452112
##  COMPUTER SYSTEMS ANALYSTS            : 59967   Median :541511
##  SOFTWARE DEVELOPERS, SYSTEMS SOFTWARE: 27135   Mean   :443246
##  COMPUTER PROGRAMMERS                 : 26681   3rd Qu.:541511
##  COMPUTER SYSTEMS ANALYST             : 14262   Max.   :928120
##  (Other)                              :270909   NA's   :6
##  PREVAILING_WAGE       PW_UNIT_OF_PAY   ENTRY_MISTAKE
##  Min.   :0.000e+00            :    57   Mode :logical
##  1st Qu.:6.463e+04   Bi-Weekly:    44   FALSE:654299
##  Median :8.012e+04   Hour     : 44899   TRUE :61
##  Mean   :8.205e+04   Month    :   285
##  3rd Qu.:9.709e+04   Week     :    96
##  Max.   :1.000e+09   Year     :608979
##  NA's   :3
##                                  EMPLOYER_NAME    H1B_DEPENDENT
##  DELOITTE CONSULTING LLP               : 16140   N   :412491
##  TATA CONSULTANCY SERVICES LIMITED     : 14604   Y   :227713
##  INFOSYS LIMITED                       : 11591   NA's: 14156
##  COGNIZANT TECHNOLOGY SOLUTIONS US CORP: 11086
##  ERNST & YOUNG U.S. LLP                :  6892
##  ACCENTURE LLP                         :  6381
##  (Other)                               :587666
```

```
##                       OCCUPATIONAL_CLASSIFICATION
##  Computer and Mathematical         :446178
##  Business and Financial Operations : 54524
##  Architecture and Engineering      : 49012
##  Management                        : 25582
##  Life, Physical, and Social Science: 23522
##  (Other)                           : 54946
##  NA's                              :   596
```

Analyze Missing Pattern

```
#check missing data pattern
library(extracat)
visna(H1B)
```



Missing Data Pattern:

By looking at missing data pattern, we tend to conclude that the quality of this dataset is good. Most variables either have no missing values or have smaller than 0.01% missing value. Variable "H1B_DEPENDENT" has most missing values, with only 2% missing value proportion. Thus, our visualization and analysis won't eroded by missing values.

Analyze Variables

Given 26 variables:

Categorical: CASE_STATUS, WAGE_UNIT_OF_PAY, MAJOR_SOC_CODE, MAJOR_NAICS_CODE, MAJOR_INDUSTRY, EMPLOYER_CITY, EMPLOYER_STATE, EMPLOYER_POSTAL_CODE, EMPLOYER_COUNTRY, SOC_CODE, SOC_NAME, NAICS_CODE, PW_UNIT_OF_PAY, ENTRY_MISTAKE, H1B_DEPENDENT, OCCUPATIONAL_CLASSIFICATION

Discrete:WAIT_TIME

Continuous: ANNUAL_WAGE, WAGE_RATE_OF_PAY_FROM, PREVAILING_WAGE

Date: CASE_SUBMITTED, DECISION_DATE

Textual Data: CASE_NUMBER, JOB_TITLE, EMPLOYER_NAME

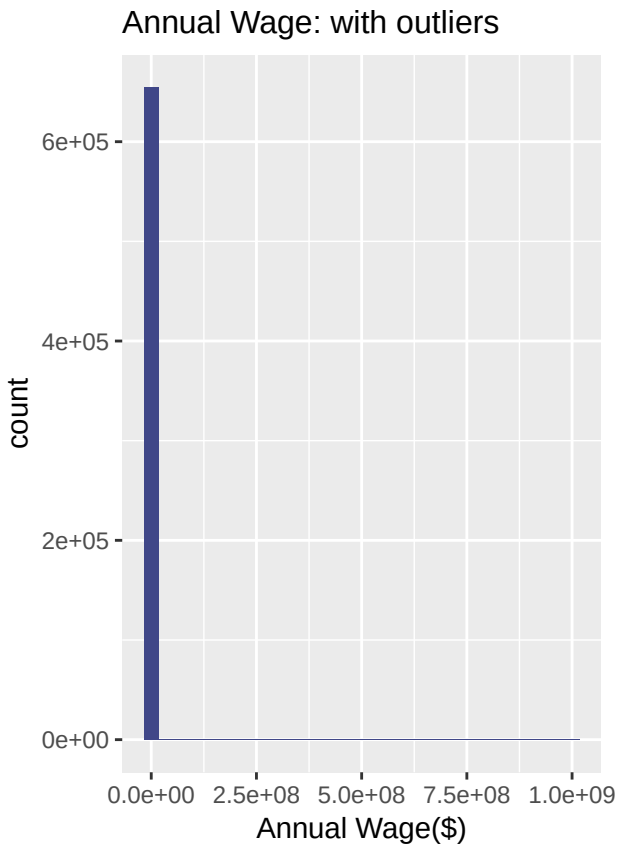
Analysis of Key Continuous Variables:

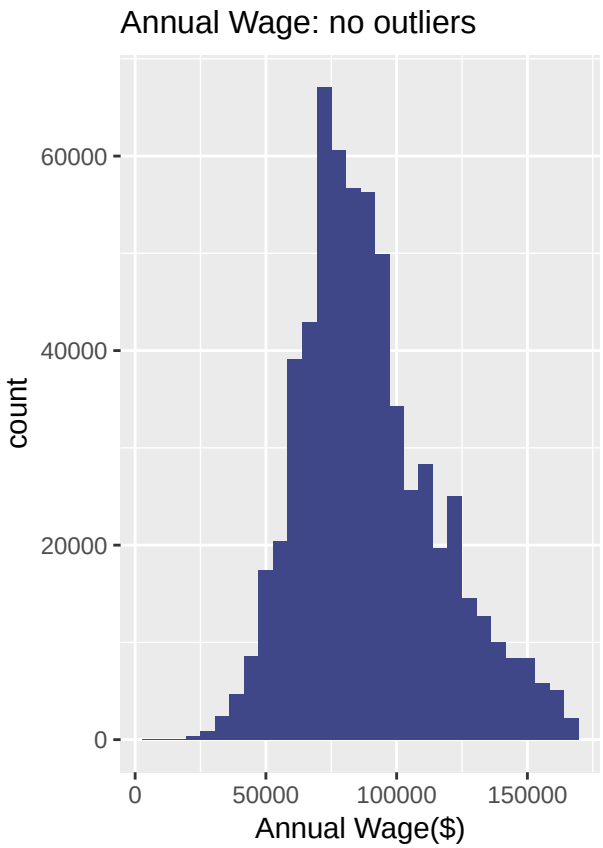
```
library(ggplot2)
library(ggpubr)
library(tidyverse)

#Variable "annual wage"
hist1 <- ggplot(data=H1B, aes(H1B$ANNUAL_WAGE)) + geom_histogram(fill="#404788FF") +
  ggtitle("Annual Wage: with outliers") + xlab("Annual Wage($)") +
  theme(plot.title = element_text(size=12))

#Drop outliers
upper <- summary(H1B$ANNUAL_WAGE)[5]+
  (summary(H1B$ANNUAL_WAGE)[5]-summary(H1B$ANNUAL_WAGE)[2])*1.5
H1B_clean <- H1B %>% filter(H1B$ANNUAL_WAGE<=upper & H1B$ANNUAL_WAGE>0)

hist2<- ggplot(data=H1B_clean, aes(ANNUAL_WAGE)) + geom_histogram(fill="#404788FF") +
  ggtitle("Annual Wage: no outliers") + xlab("Annual Wage($)") +
  theme(plot.title = element_text(size=12))
```

Analysis of Key Discrete Variables:

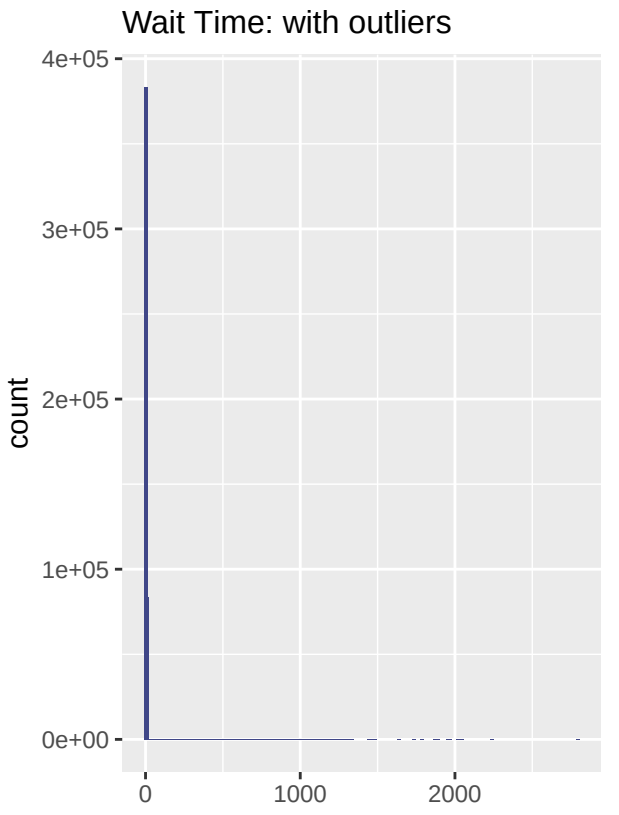
```
#Variable "wait time"
H1B_waittime <- H1B %>% group_by(WAIT_TIME) %>% summarize(count=n())
hist3<-ggplot(data=H1B_waittime, aes(x=WAIT_TIME, y=count)) +
  geom_bar(stat="identity", fill="#404788FF",width = 20) + ggtitle("Wait Time: with outliers") +
  xlab("")+ theme(plot.title = element_text(size=12))

#Drop outliers
H1B_clean_waittime <- H1B %>% filter(H1B$WAIT_TIME<=20 & H1B$WAIT_TIME>0)

H1B_waittime2 <- H1B_clean_waittime %>% group_by(WAIT_TIME) %>% summarize(count=n())
hist4<-ggplot(data=H1B_waittime2, aes(x=WAIT_TIME, y=count)) +
  geom_bar(stat="identity", fill="#404788FF") + ggtitle("Wait Time: no outliers") + xlab("")+
  theme(plot.title = element_text(size=12))

ggarrange(hist1,hist2, nrow=1, ncol = 2)
```

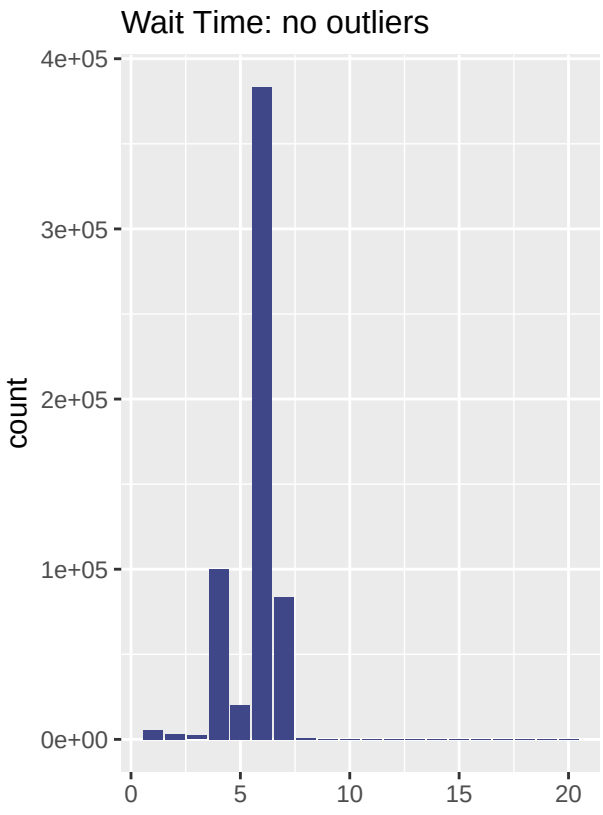
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```






```
ggarrange(hist3,hist4, nrow=1, ncol = 2)
```

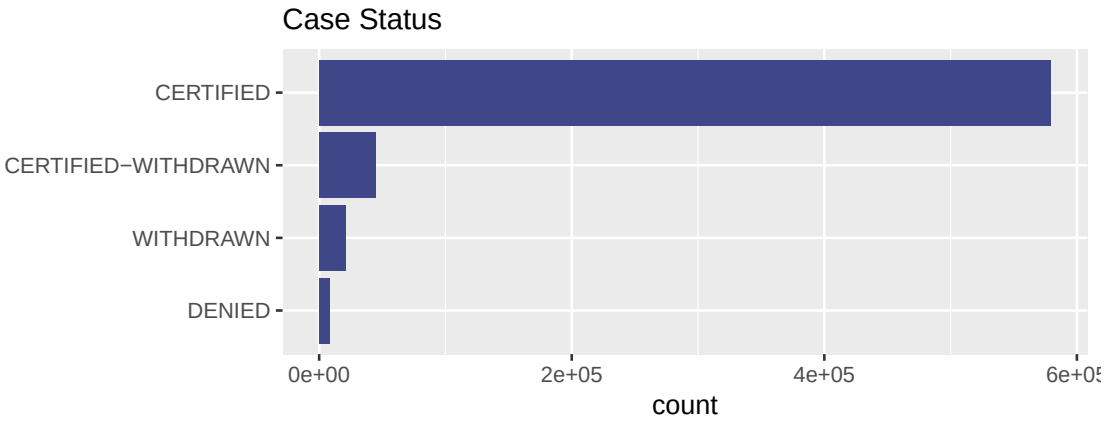


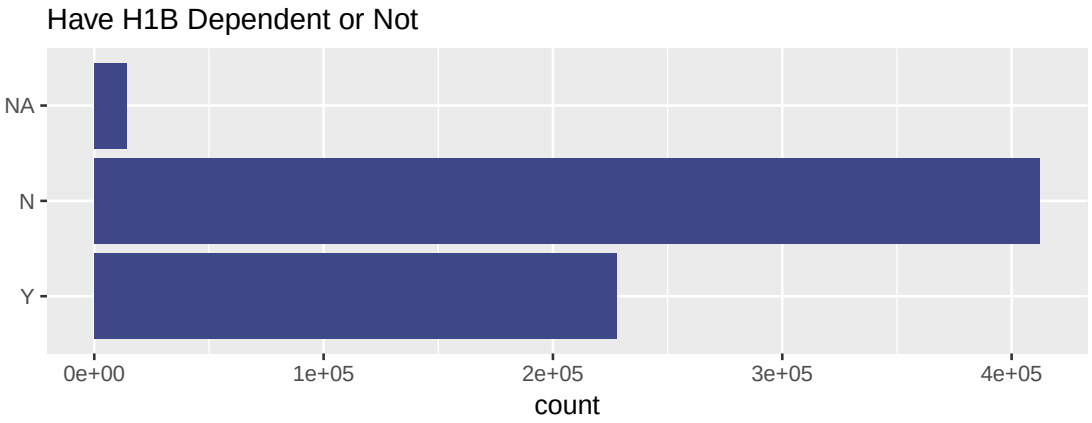


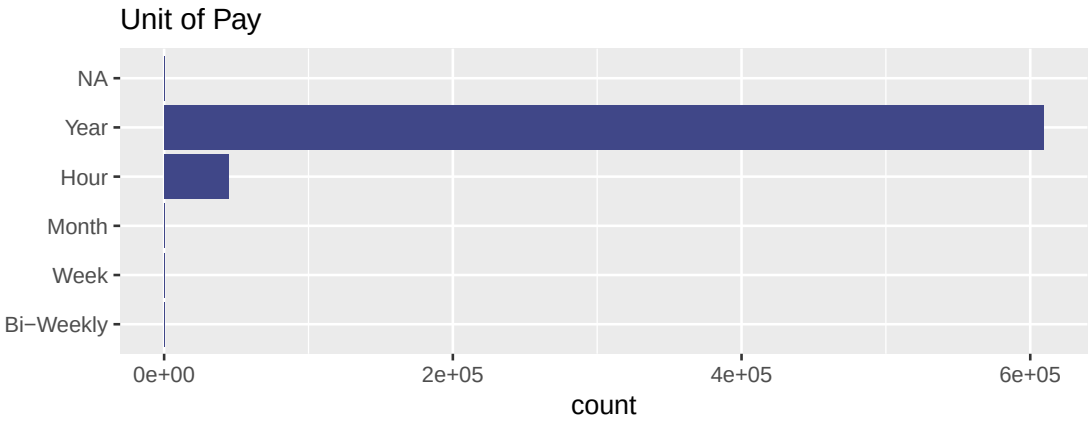
Analysis of Key Categorical Variables:

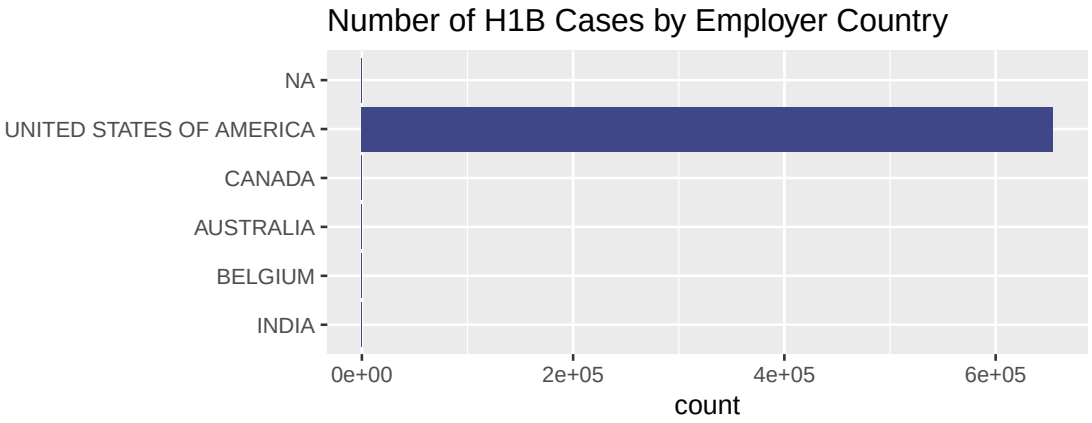
```
H1B_status <- H1B %>% group_by(CASE_STATUS) %>% summarize(count=n())
bar1 <- ggplot(data=H1B_status, aes(reorder(x=CASE_STATUS), y=count),label=Frequency) +
  geom_bar(stat="identity", fill="#404788FF") + ggtitle("Case Status") + xlab("")+
  theme(plot.title = element_text(size=12))+coord_flip()

H1B_dependent <- H1B %>% group_by(H1B_DEPENDENT) %>% summarize(count=n())
bar2 <-ggplot(data=H1B_dependent, aes(reorder(x=H1B_DEPENDENT, count), y=count)) +
  geom_bar(stat="identity", fill="#404788FF") + ggtitle("Have H1B Dependent or Not") + xlab("")+
  theme(plot.title = element_text(size=12))+coord_flip()

H1B_unitpay <- H1B %>% group_by(WAGE_UNIT_OF_PAY) %>% summarize(count=n())
bar3 <- ggplot(data=H1B_unitpay, aes(reorder(x=WAGE_UNIT_OF_PAY, count), y=count)) +
  geom_bar(stat="identity", fill="#404788FF") + ggtitle("Unit of Pay") + xlab("")+
  theme(plot.title = element_text(size=12))+coord_flip()

H1B_country <- H1B %>% group_by(EMPLOYER_COUNTRY) %>% summarize(count=n())
bar4 <- ggplot(data=H1B_country, aes(reorder(x=EMPLOYER_COUNTRY, count), y=count)) +
  geom_bar(stat="identity", fill="#404788FF") + ggtitle("Number of H1B Cases by Employer Country")  +
  coord_flip() + xlab(" ")

H1B_industry <- H1B %>% group_by(MAJOR_INDUSTRY) %>% summarize(count=n())
bar5 <-ggplot(data=H1B_industry, aes(reorder(x=MAJOR_INDUSTRY, count), y=count)) +
  geom_bar(stat="identity", fill="#404788FF") + ggtitle("Major Industry")  +
  coord_flip() + xlab(" ")

H1B_occupation <- H1B %>% group_by(OCCUPATIONAL_CLASSIFICATION) %>% summarize(count=n())
bar6 <-ggplot(data=H1B_occupation, aes(reorder(x=OCCUPATIONAL_CLASSIFICATION, count), y=count)) +
  geom_bar(stat="identity", fill="#404788FF") + ggtitle("Occupational Classification")  +
  coord_flip() + xlab(" ")

H1B_state <- H1B %>% group_by(EMPLOYER_STATE) %>% summarize(count=n())
bar7 <-ggplot(data=H1B_state, aes(reorder(x=EMPLOYER_STATE, count), y=count)) +
  geom_bar(stat="identity", fill="#404788FF") + ggtitle("Number of H1B Cases by Employer State")  +
  coord_flip() + xlab(" ")+ theme(axis.text.y=element_text(size = 6)) +
  geom_text(aes(label = count), nudge_y = 2)
```

```
ggarrange(bar1, bar2, bar3, bar4, ncol = 1, nrow = 4)
```

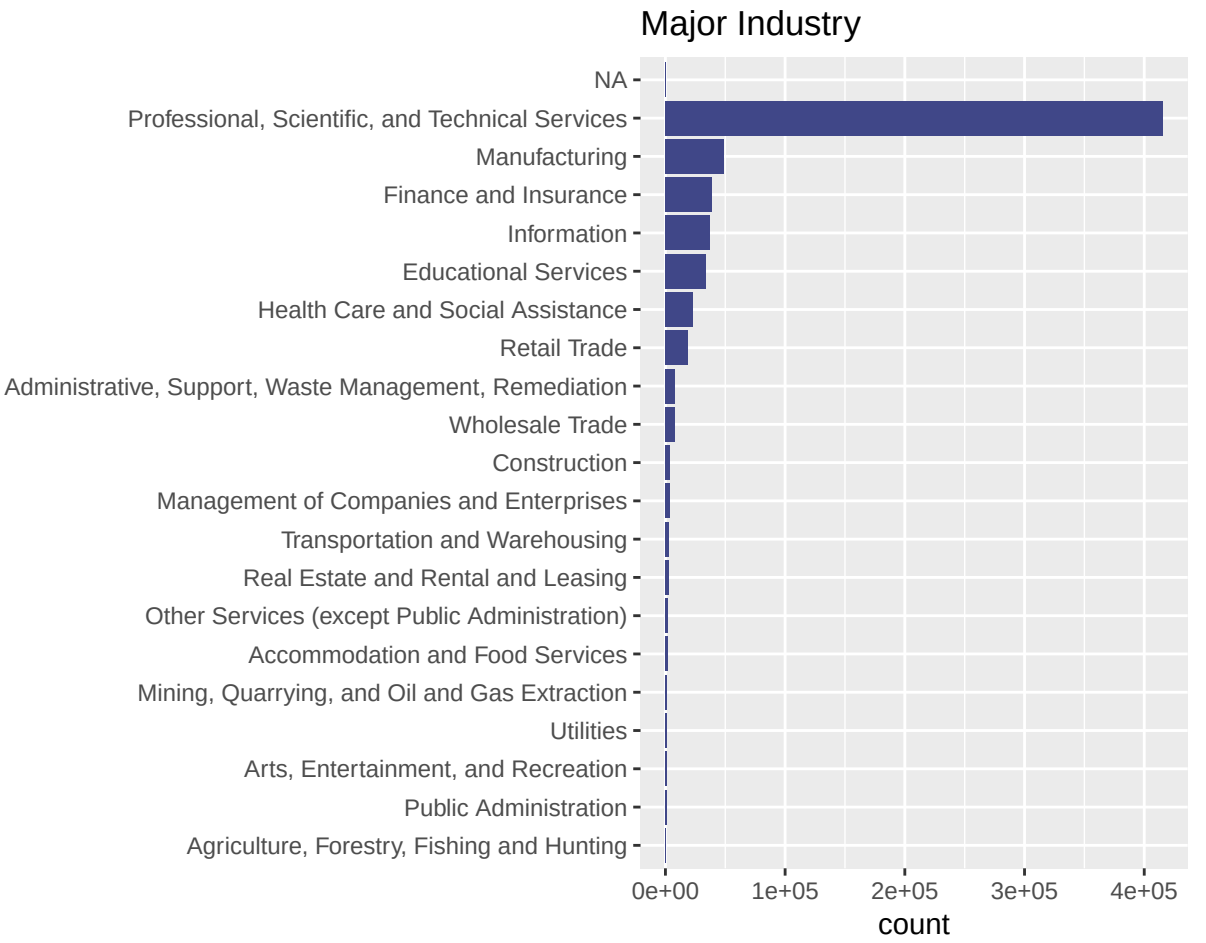

Case Status
CERTIFIED
CERTIFIED−WITHDRAWN
WITHDRAWN
DENIED
0e+00
2e+05
4e+05
6e+05
count


Have H1B Dependent or Not
NA
N
Y
0e+00
1e+05
2e+05
3e+05
4e+05
count

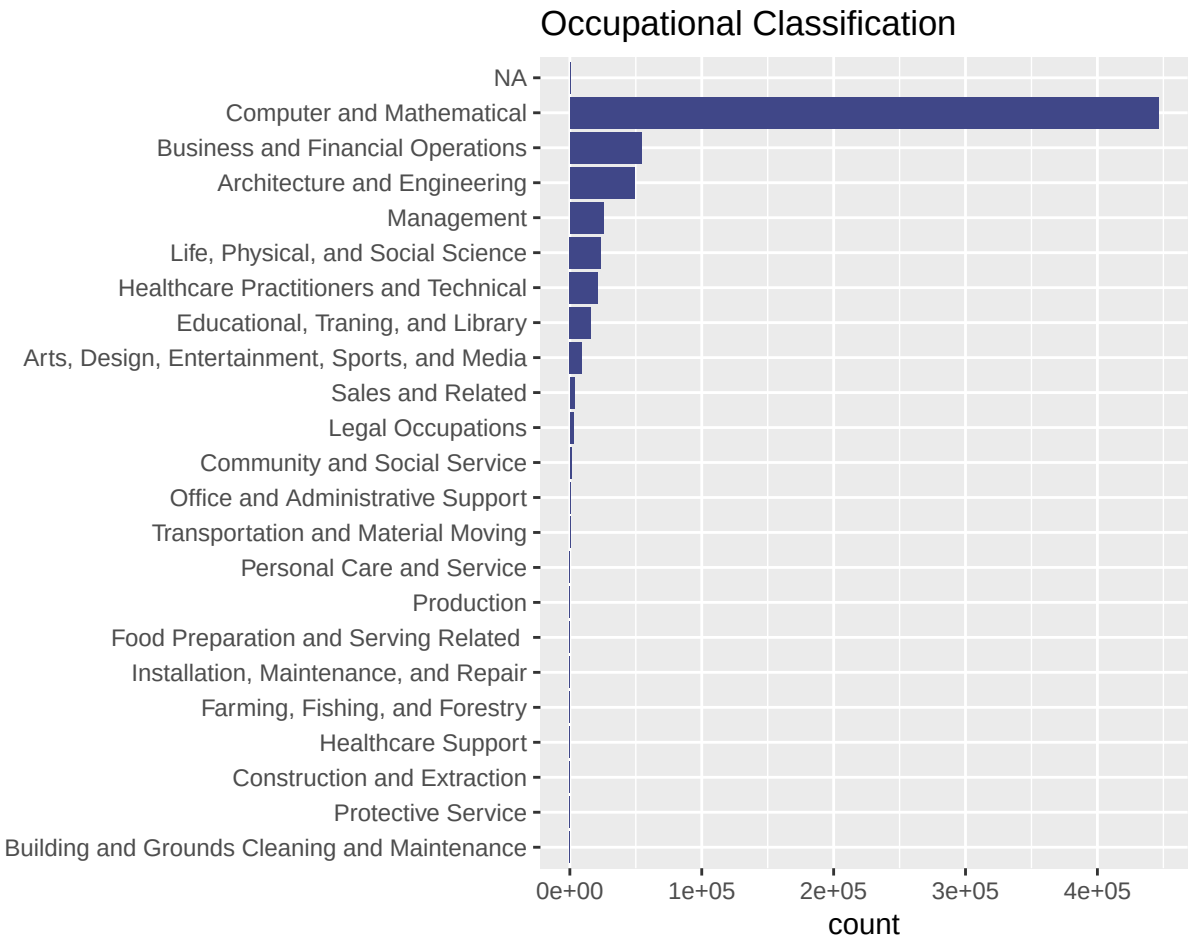

Unit of Pay
NA
Year
Hour
Month
Week
Bi−Weekly
0e+00
2e+05
4e+05
6e+05
count


Number of H1B Cases by Employer Country
NA
UNITED STATES OF AMERICA
CANADA
AUSTRALIA
BELGIUM
INDIA
0e+00
2e+05
4e+05
6e+05
count

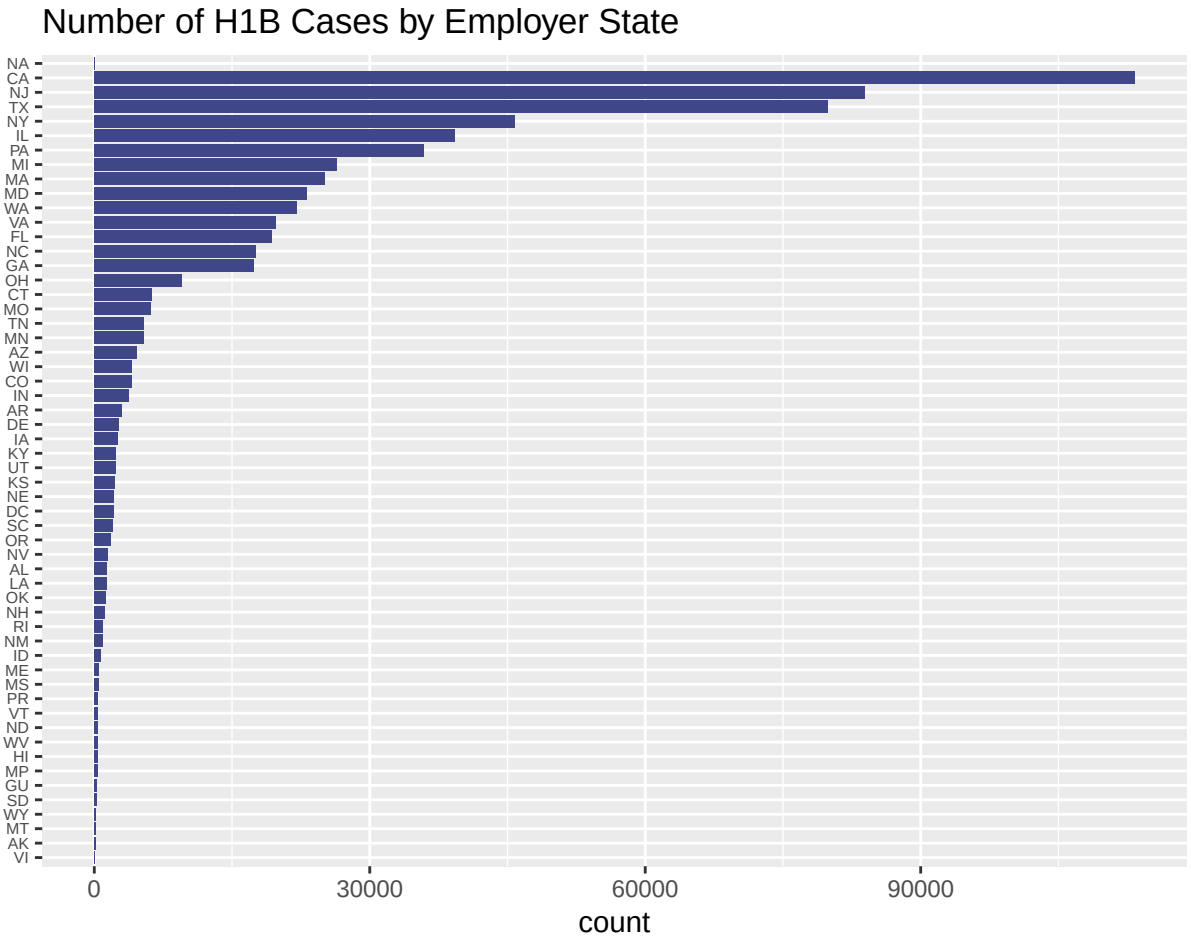
```
bar5
```


Major Industry
NA
Professional, Scientific, and Technical Services
Manufacturing
Finance and Insurance
Information
Educational Services
Health Care and Social Assistance
Retail Trade
Administrative, Support, Waste Management, Remediation
Wholesale Trade
Construction
Management of Companies and Enterprises
Transportation and Warehousing
Real Estate and Rental and Leasing
Other Services (except Public Administration)
Accommodation and Food Services
Mining, Quarrying, and Oil and Gas Extraction
Utilities
Arts, Entertainment, and Recreation
Public Administration
Agriculture, Forestry, Fishing and Hunting
0e+00
1e+05
2e+05
3e+05
4e+05
count

```
bar6
```


Occupational Classification
NA
Computer and Mathematical
Business and Financial Operations
Architecture and Engineering
Management
Life, Physical, and Social Science
Healthcare Practitioners and Technical
Educational, Traning, and Library
Arts, Design, Entertainment, Sports, and Media
Sales and Related
Legal Occupations
Community and Social Service
Office and Administrative Support
Transportation and Material Moving
Personal Care and Service
Production
Food Preparation and Serving Related
Installation, Maintenance, and Repair
Farming, Fishing, and Forestry
Healthcare Support
Construction and Extraction
Protective Service
Building and Grounds Cleaning and Maintenance
0e+00
1e+05
2e+05
3e+05
4e+05
count

```
bar7
```



Key Variables of Interest:

- Variable "ANNUAL_WAGE" is influenced by large value outliers and it follows normal distribution after dropping outliers.
- Variable "WAIT_TIME" is also influenced by large value outliers. After dropping variables, we observe that most of wait time are of H1B application are around 6 days.
- Variable "CASE_STATUS": most cases are certified, which is siginificantly greater than other status.The second most case statu is certified-withdrawn.
- Variable "H1B_DEPENDENT": the number of applicants who don't have H1B dependents are approximately twice as those who have H1B dependents
- Variable "WAGE_UNIT_OF_PAY": most of wages are in the unit pay of year, with only a few in hour.
- Variable "EMPLOYER_COUNTRY": most of employer country is the United States.
- Variable "MAJOR_INDUSTRY": Professional, Scientific, and Technical Services Industry has most H1B cases, much more than other industries. Manufacturing is in the second position, and Finance and Insurance is in the thrid position.
- Variable "OCCUPATIONAL_CLASSIFICATION": Computer and Mathematical occupation has most H1B cases, far more than other occupational calssification.
- Variable "EMPLOYER_STATE": the top 3 U.S. states with most H1B cases are California, New Jersey, Texas.