May 2017 National Occupational Employment and Wage Estimates United States

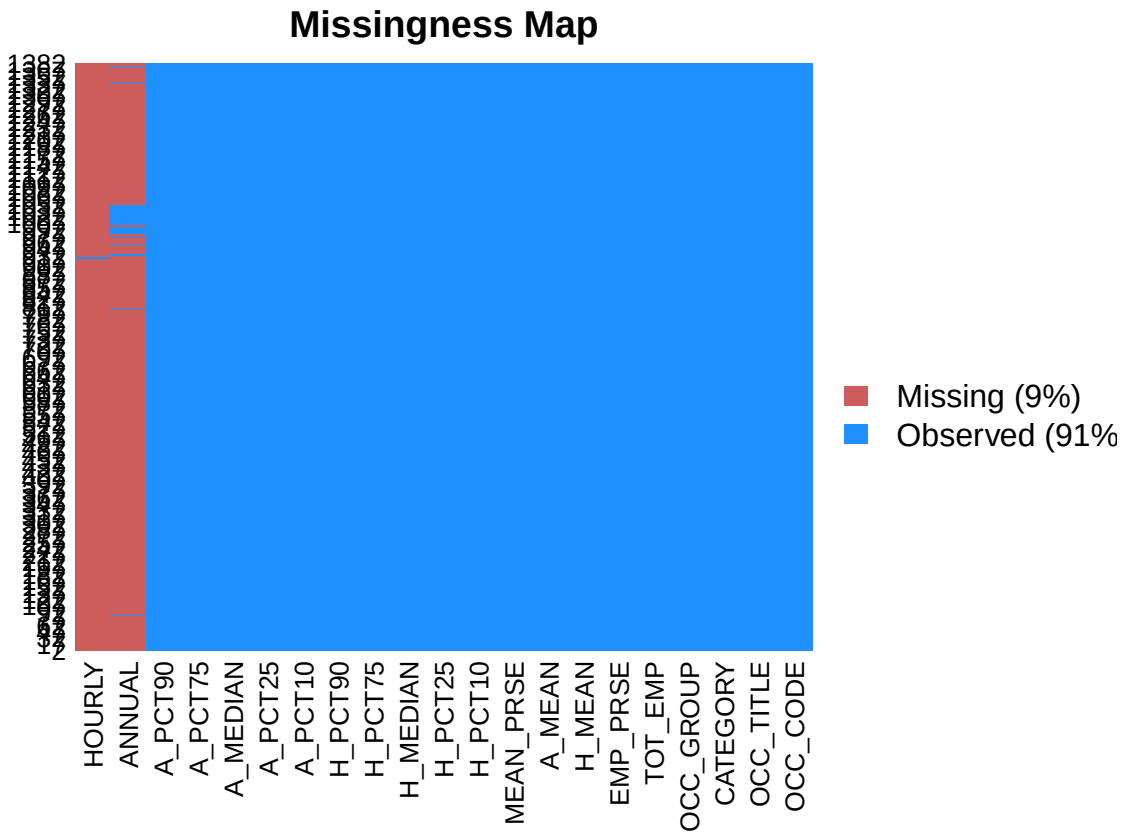
7/3/2018

```
library(scales)
library(readxl)
library(Amelia)
library(psych)
library(knitr)
library(ggplot2)
library(ggalluvial)

#loading file
M2017 <- read_excel('national_M2017_dl.xlsx')

#file structure
kable(describe(M2017))
```

	vars	n	mean	sd	median	trimmed	mad	min	
OCC_CODE*	1	1382	NaN	NA	NA	NaN	NA	Inf	
OCC_TITLE*	2	1382	NaN	NA	NA	NaN	NA	Inf	
CATEGORY*	3	1382	NaN	NA	NA	NaN	NA	Inf	
OCC_GROUP*	4	1382	NaN	NA	NA	NaN	NA	Inf	
TOT_EMP	5	1382	5.157351e+05	4.019189e+06	73940.000	1.481169e+05	92988.672000	390.00	1
EMP_PRSE	6	1382	2.976845e+00	3.158043e+00	2.000	2.395208e+00	1.630860	0.00	
H_MEAN*	7	1382	2.778635e+01	1.539189e+01	23.600	2.534661e+01	10.348548	10.21	
A_MEAN*	8	1382	5.898305e+04	3.201864e+04	50200.000	5.422345e+04	22891.344000	21230.00	
MEAN_PRSE	9	1382	1.090666e+00	1.049495e+00	0.800	9.092224e-01	0.593040	0.10	
H_PCT10*	10	1382	1.518747e+01	6.824408e+00	12.855	1.403958e+01	4.588647	8.07	
H_PCT25*	11	1382	1.942431e+01	1.009474e+01	16.370	1.779759e+01	7.042350	8.74	
H_MEDIAN*	12	1382	2.510761e+01	1.284436e+01	21.580	2.320403e+01	9.829638	9.53	
H_PCT75*	13	1382	3.239511e+01	1.591192e+01	28.590	3.029697e+01	13.002402	11.22	
H_PCT90*	14	1382	3.977486e+01	1.833543e+01	36.230	3.770195e+01	16.486512	12.85	
A_PCT10*	15	1382	3.193047e+04	1.403603e+04	27465.000	2.969328e+04	10385.613000	16790.00	
A_PCT25*	16	1382	4.096450e+04	2.075907e+04	35090.000	3.780749e+04	15908.298000	18180.00	
A_MEDIAN*	17	1382	5.315746e+04	2.665797e+04	46670.000	4.947039e+04	21334.614000	19820.00	
A_PCT75*	18	1382	6.896380e+04	3.353071e+04	61175.000	6.486060e+04	28881.048000	23340.00	
A_PCT90*	19	1382	8.500178e+04	3.935388e+04	76930.000	8.080585e+04	36160.614000	26740.00	
ANNUAL*	20	82	NaN	NA	NA	NaN	NA	Inf	
HOURLY*	21	6	NaN	NA	NA	NaN	NA	Inf	

Before the analysis, I categorized all the occupation using https://www.cityofmadison.com/dcr/documents/EEO-1JobCat.pdf to categorize the occupation into relevant category.

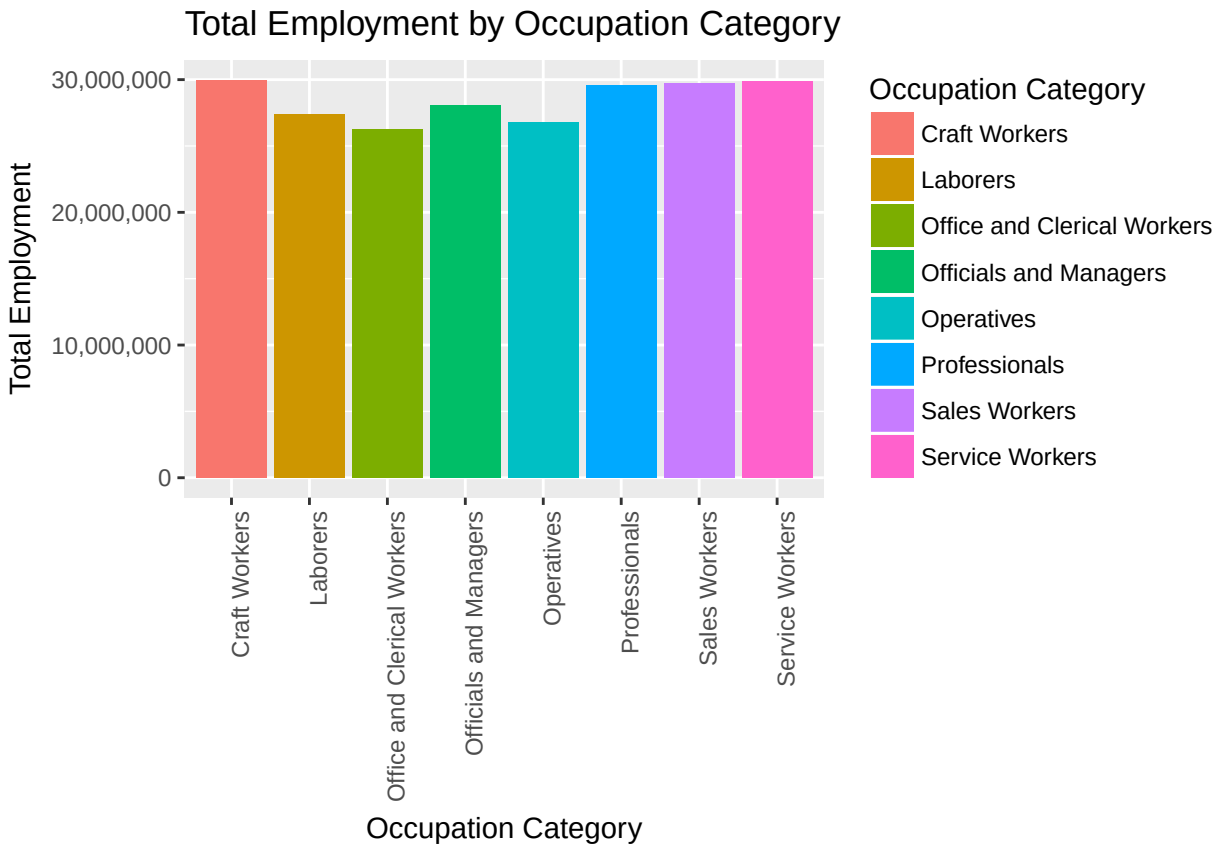
```
#checking missing values
missmap(M2017)
```



Due to large amount of missing value in "HOURLY" and "ANNUAL" columns, it is suggested that to remove these two columns.

```
# using mean
M2017_1 <- M2017[,c(2,3,4,5,7,8)]

#remove 1st row
M2017_1 <- M2017_1[-1,]


#convert into relevant data type
M2017_1$OCC_TITLE <- as.character(M2017_1$OCC_TITLE)
M2017_1$OCC_GROUP <- as.factor(M2017_1$OCC_GROUP)
M2017_1$CATEGORY <- as.factor(M2017_1$CATEGORY)
M2017_1$TOT_EMP <- as.numeric(M2017_1$TOT_EMP)
M2017_1$H_MEAN <- as.numeric(M2017_1$H_MEAN)
M2017_1$A_MEAN <- as.numeric(M2017_1$A_MEAN)

#remove missing value
M2017_1.na <- na.omit(M2017_1)

#total employment by group
ggplot(M2017_1.na, aes(x = OCC_GROUP, y = TOT_EMP, fill = OCC_GROUP)) +
  geom_bar(stat = 'identity') +
  scale_y_continuous(limits = c(-0,30000000), labels = comma) +
  labs(title = "Total Employment by Occupation Group", x = 'Occupation Group', y = "Total Employment") +
  guides(fill = guide_legend(title = "Occupation Group"))
```


```
#total employment by category
ggplot(M2017_1.na, aes(x = CATEGORY, y = TOT_EMP, fill = CATEGORY)) +
  geom_bar(stat = 'identity') +
  scale_y_continuous(limits = c(-0,30000000), labels = comma) +
  labs(title = "Total Employment by Occupation Category", x = 'Occupation Category', y = "Total Employme
  guides(fill = guide_legend(title = "Occupation Category")) +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```




```
#hourly wage
summary(M2017_1.na$H_MEAN)
```

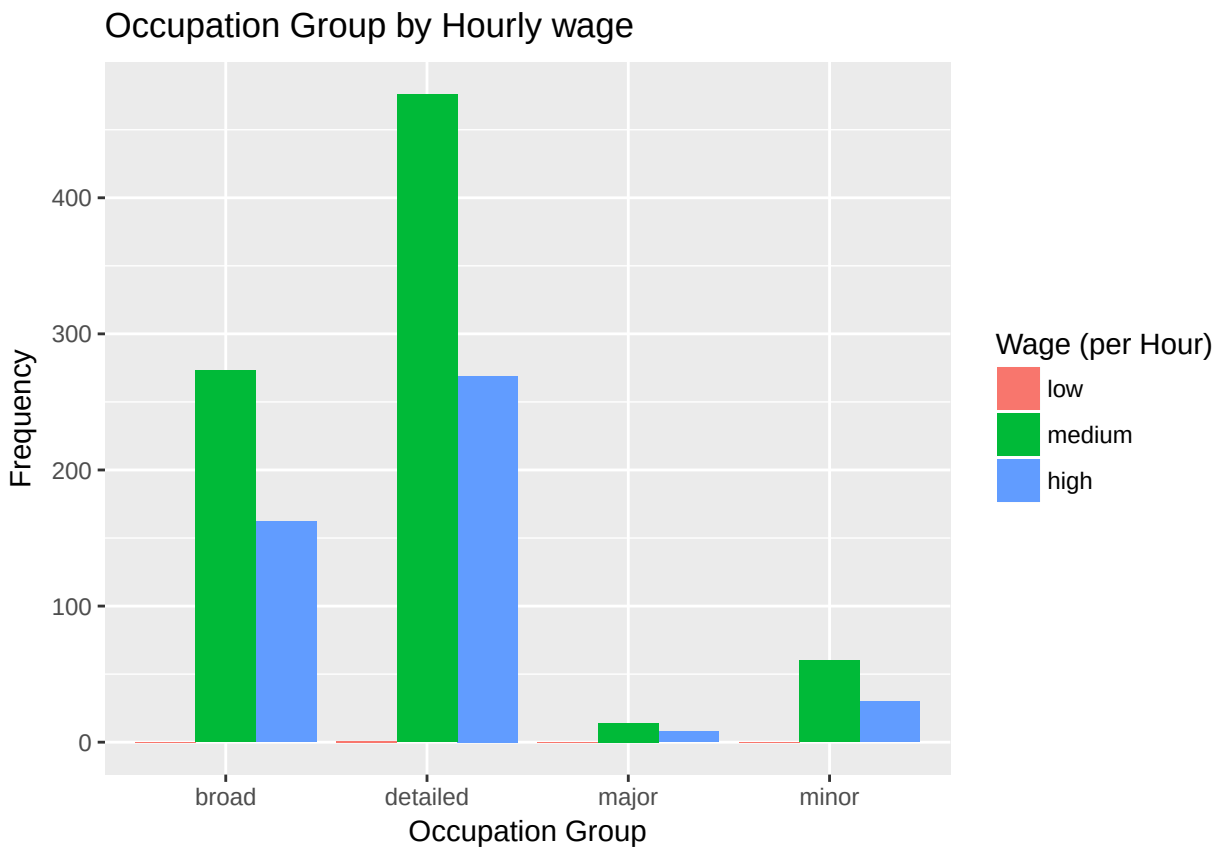
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.21   17.43   23.60   27.79   33.26  127.88
```

```
M2017_1.na$Hour_wage <- cut(M2017_1.na$H_MEAN,
                            breaks = c(-Inf, 10.21, 27.74, Inf),
                            labels = c("low", "medium", "high"))

#annual income
M2017_1.na$Income <- cut(M2017_1.na$A_MEAN,
                            breaks = c(-Inf, 49999, 99999, Inf),
                            labels = c("low", "medium", "high"))
```

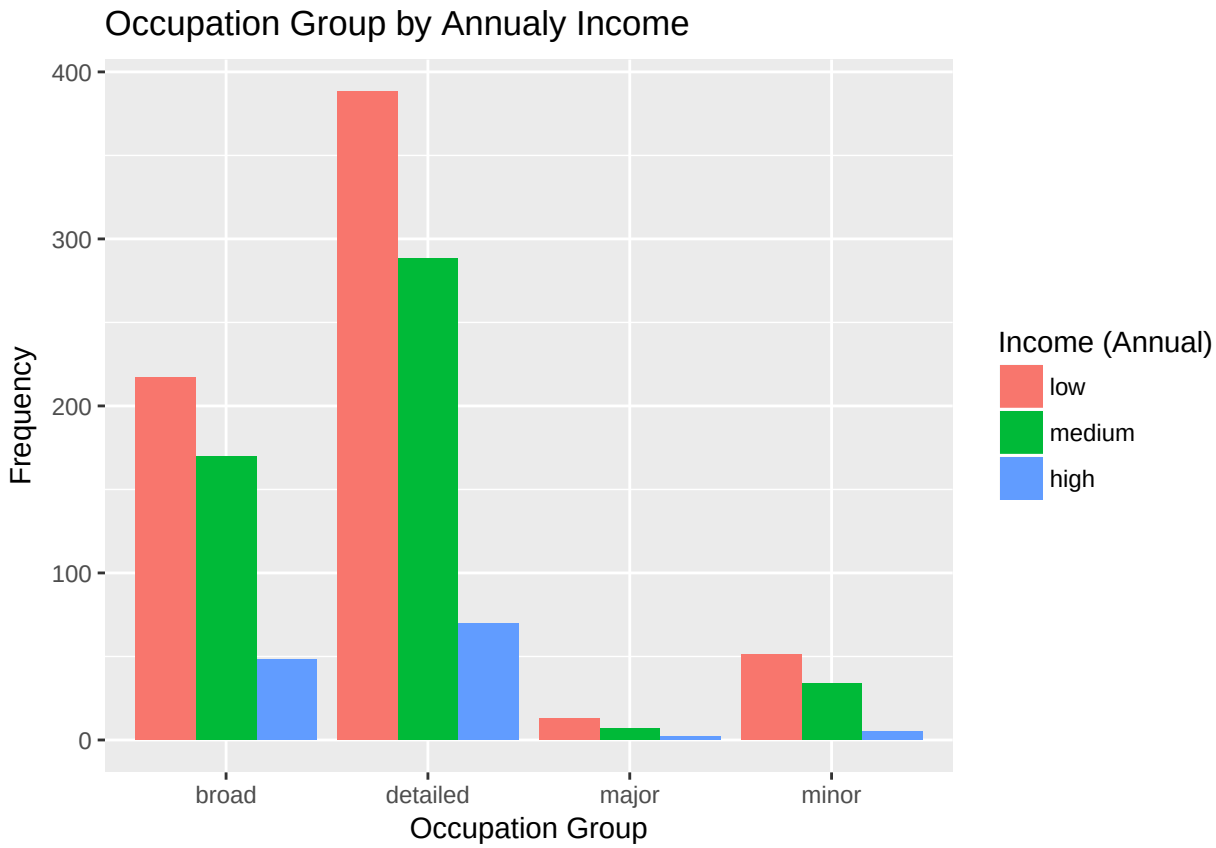
Due to lack of information for the hourly wage in USA, I decided to use the mean instead. I categorized 10.21 as low hourly wages, $>$ 27.74 $<=$ as medium hourly wage, and anything above that is consider as high. To categorize the annual income, I used Middle Class, Income Are You in the Middle Class? published by the website the balance https://www.thebalance.com/definition-of-middle-class-income-4126870.

```
#catgorize the data
df <- table(M2017_1.na$OCC_GROUP, M2017_1.na$Hour_wage)
df1 <- table(M2017_1.na$OCC_GROUP, M2017_1.na$Income)
df2 <- table(M2017_1.na$CATEGORY, M2017_1.na$Hour_wage)
df3 <-  table(M2017_1.na$CATEGORY, M2017_1.na$Income)

df <- as.data.frame(df)
df1 <- as.data.frame(df1)
```

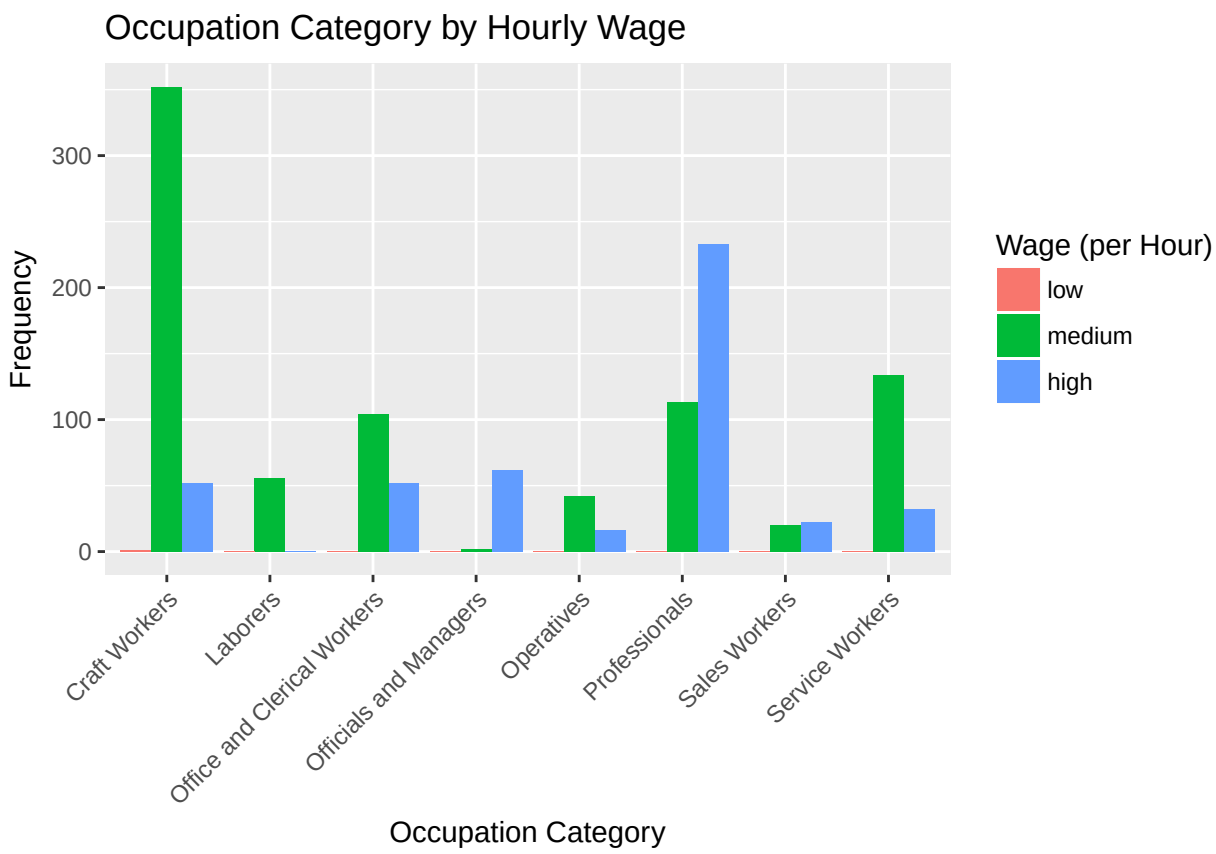
```r
df2 <- as.data.frame(df2)
df3 <- as.data.frame(df3)

colnames(df) <- c("Group", "Wage (per Hour)", "Freq")
colnames(df1) <- c("Category", "Income (Annual)", "Freq")
colnames(df2) <- c("Group", "Wage (per Hour)", "Freq")
colnames(df3) <- c("Category", "Income (Annual)", "Freq")

#plot
ggplot(df, aes(x = Group, y = Freq, fill = `Wage (per Hour)`)) +
  geom_bar(stat = "identity", position = 'dodge') +
  labs(title = "Occupation Group by Hourly wage", x = 'Occupation Group', y = "Frequency")
```



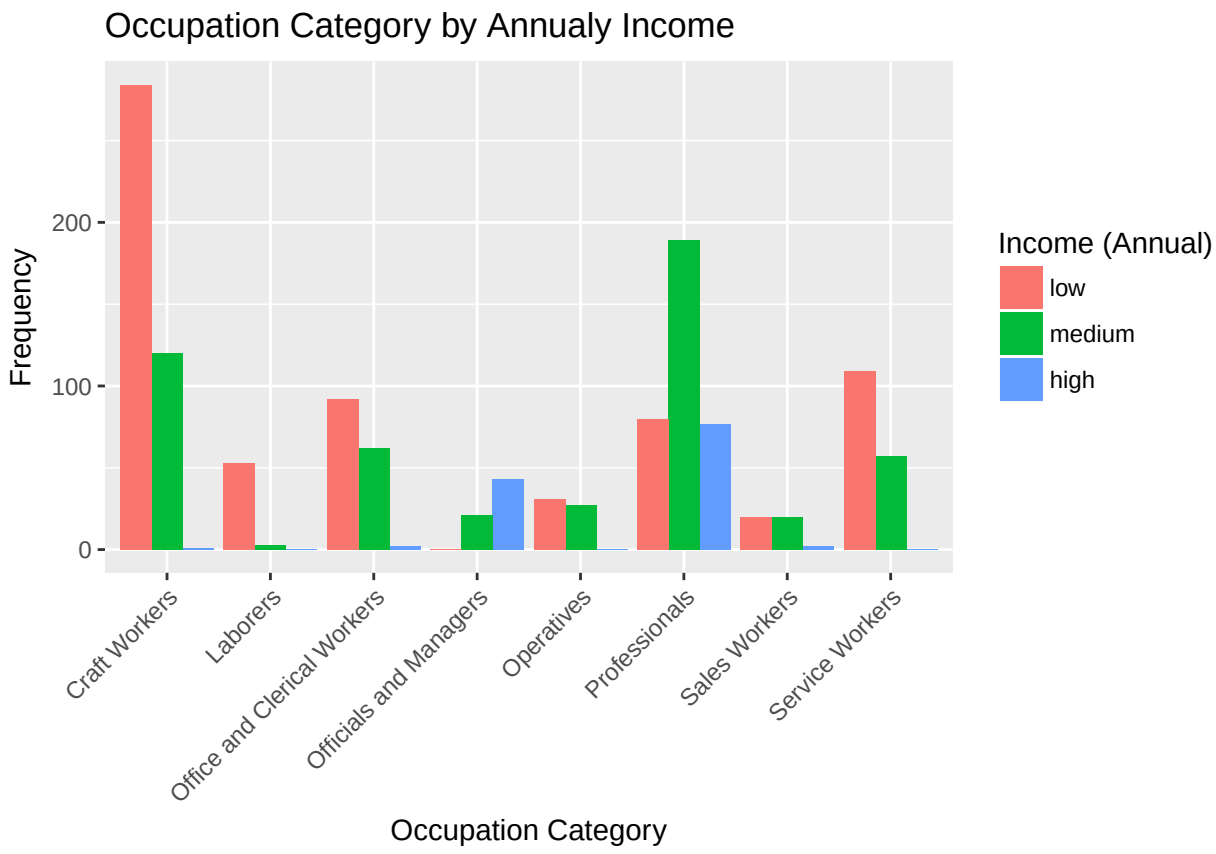

```r
ggplot(df1, aes(x = Category, y = Freq, fill = `Income (Annual)`)) +
  geom_bar(stat = "identity", position = 'dodge') +
  labs(title = "Occupation Group by Annualy Income", x = 'Occupation Group', y = "Frequency")
```



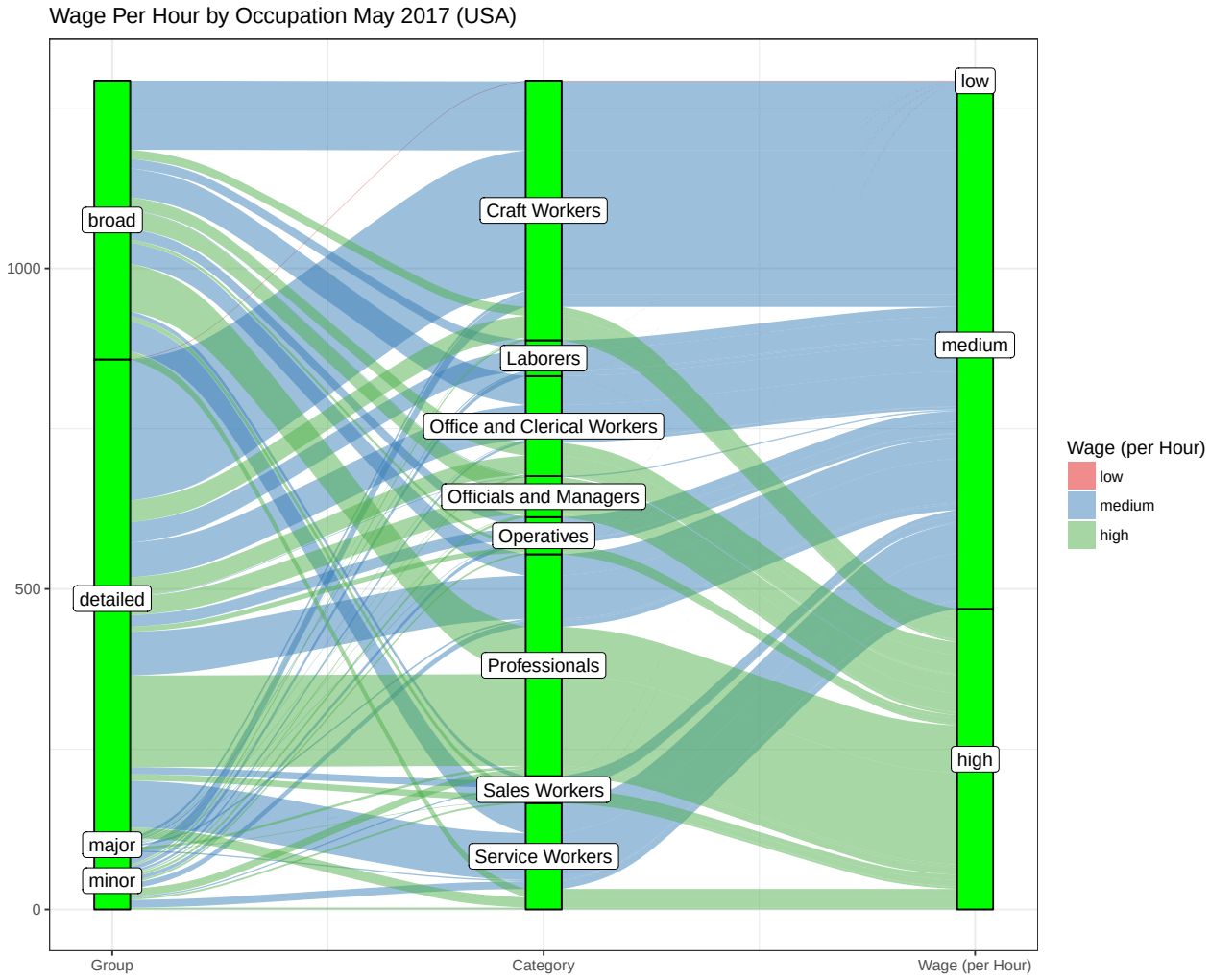

```
ggplot(df2, aes(x = Group, y = Freq, fill = `Wage (per Hour)`)) +
  geom_bar(stat = "identity", position = 'dodge') +
  labs(title = "Occupation Category by Hourly Wage", x = 'Occupation Category', y = "Frequency") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



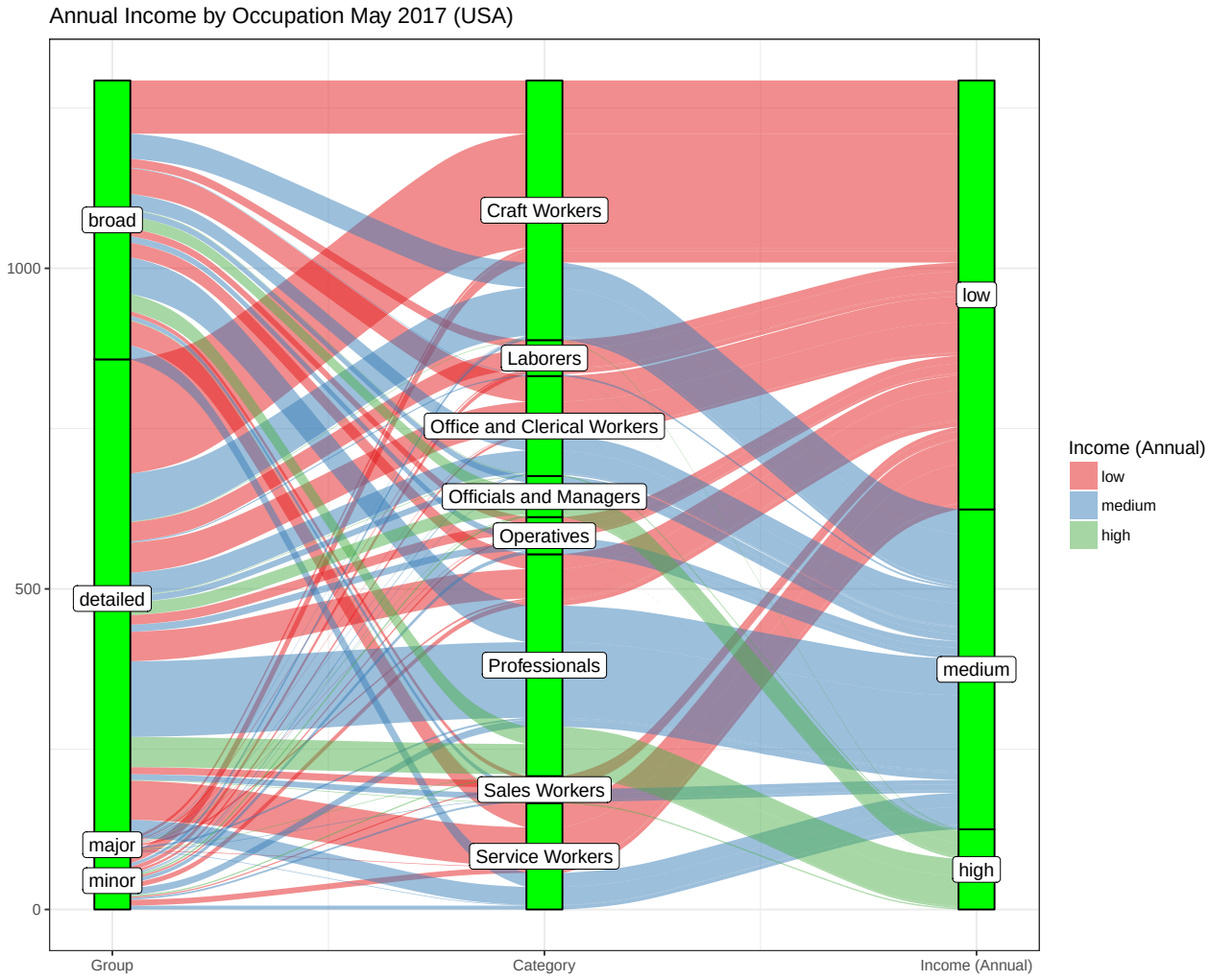

```
ggplot(df3, aes(x = Category, y = Freq, fill = `Income (Annual)`)) +
  geom_bar(stat = "identity", position = 'dodge') +
  labs(title = "Occupation Category by Annualy Income", x = 'Occupation Category', y = "Frequency") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```




```
#sankey diagram for Wage per hour
df4 <- table(M2017_1.na$OCC_GROUP, M2017_1.na$CATEGORY, M2017_1.na$Hour_wage)
df4 <- as.data.frame(df4)
colnames(df4) <- c("Group", "Category", "Wage (per Hour)", "Freq")

ggplot(df4,
       aes(weight = Freq, axis1 = Group, axis2 = Category, axis3 = `Wage (per Hour)`)) +
  geom_alluvium(aes(fill = `Wage (per Hour)`), width = 1/12) +
  geom_stratum(width = 1/12, fill = "green", color = "black") +
  geom_label(stat = "stratum", label.strata = TRUE) +
  scale_x_continuous(breaks = 1:3, labels = c("Group", "Category", "Wage (per Hour)")) +
  scale_fill_brewer(type = "qual", palette = "Set1") +
  ggtitle("Wage Per Hour by Occupation May 2017 (USA)") +
  theme_bw()
```




```
#sankey diagram for annual income
df5 <- table(M2017_1.na$OCC_GROUP, M2017_1.na$CATEGORY, M2017_1.na$Income)
df5 <- as.data.frame(df5)
colnames(df5) <- c("Group", "Category", "Income (Annual)", "Freq")

ggplot(df5,
       aes(weight = Freq, axis1 = Group, axis2 = Category, axis3 = `Income (Annual)`)) +
  geom_alluvium(aes(fill = `Income (Annual)`), width = 1/12) +
  geom_stratum(width = 1/12, fill = "green", color = "black") +
  geom_label(stat = "stratum", label.strata = TRUE) +
  scale_x_continuous(breaks = 1:3, labels = c("Group", "Category", "Income (Annual)")) +
  scale_fill_brewer(type = "qual", palette = "Set1") +
  ggtitle("Annual Income by Occupation May 2017 (USA)") +
  theme_bw()
```


Annual Income by Occupation May 2017 (USA)
1000
500
0
broad
detailed
major
minor
Craft Workers
Laborers
Office and Clerical Workers
Officials and Managers
Operatives
Professionals
Sales Workers
Service Workers
low
medium
high
Group
Category
Income (Annual)
Income (Annual)
low
medium
high