Behavioral Data Science Week 3 - Exploritory Data Analysis

Jonathan Gragg

9/12/2021

loading in relevant Libraries

```
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------------------------

## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.3     v dplyr   1.0.2
## v tidyr   1.1.1     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0

## -- Conflicts ------------------------------------------------------------------------------------
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ggplot2)
library(corrplot)
```

```
## Warning: package 'corrplot' was built under R version 4.0.3
```

```
## corrplot 0.84 loaded
```

```
library(stringr)
```

loading in the data

```
data <- read.csv("survey_results_public.csv")
glimpse(data)
```

```
## Rows: 51,392
## Columns: 154
## $ Respondent                    <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11,...
## $ Professional                  <chr> "Student", "Student", "Professiona...
## $ ProgramHobby                  <chr> "Yes, both", "Yes, both", "Yes, bo...
## $ Country                       <chr> "United States", "United Kingdom",...
## $ University                    <chr> "No", "Yes, full-time", "No", "No"...
## $ EmploymentStatus              <chr> "Not employed, and not looking for...
## $ FormalEducation               <chr> "Secondary school", "Some college/...
```

```
## $ MajorUndergrad                <chr> NA, "Computer science or software ...
## $ HomeRemote                    <chr> NA, "More than half, but not all, ...
## $ CompanySize                   <chr> NA, "20 to 99 employees", "10,000 ...
## $ CompanyType                   <chr> NA, "Privately-held limited compan...
## $ YearsProgram                  <chr> "2 to 3 years", "9 to 10 years", "...
## $ YearsCodedJob                 <chr> NA, NA, "20 or more years", "9 to ...
## $ YearsCodedJobPast             <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ DeveloperType                 <chr> NA, NA, "Other", NA, "Mobile devel...
## $ WebDeveloperType              <chr> NA, NA, NA, NA, NA, NA, NA, "Full ...
## $ MobileDeveloperType           <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ NonDeveloperType              <chr> NA, NA, NA, "Data scientist", NA, ...
## $ CareerSatisfaction            <int> NA, NA, 8, 6, 6, NA, 7, 7, 6, 6, 8...
## $ JobSatisfaction               <int> NA, NA, 9, 3, 8, NA, 6, 7, 6, 8, 9...
## $ ExCoderReturn                 <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ ExCoderNotForMe               <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ ExCoderBalance                <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ ExCoder10Years                <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ ExCoderBelonged               <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ ExCoderSkills                 <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ ExCoderWillNotCode            <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ ExCoderActive                 <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ PronounceGIF                  <chr> "With a soft \"g,\" like \"jiff\""...
## $ ProblemSolving                <chr> "Strongly agree", NA, "Strongly ag...
## $ BuildingThings                <chr> "Strongly agree", NA, "Strongly ag...
## $ LearningNewTech               <chr> "Agree", NA, "Strongly agree", "St...
## $ BoringDetails                 <chr> "Disagree", NA, "Somewhat agree", ...
## $ JobSecurity                   <chr> "Strongly agree", NA, "Agree", "So...
## $ DiversityImportant            <chr> "Agree", NA, "Strongly agree", "Ag...
## $ AnnoyingUI                    <chr> "Agree", NA, "Agree", "Agree", NA,...
## $ FriendsDevelopers             <chr> "Disagree", NA, "Somewhat agree", ...
## $ RightWrongWay                 <chr> "Somewhat agree", NA, "Disagree", ...
## $ UnderstandComputers           <chr> "Disagree", NA, "Disagree", "Stron...
## $ SeriousWork                   <chr> "Strongly agree", NA, "Agree", "St...
## $ InvestTimeTools               <chr> "Strongly agree", NA, "Somewhat ag...
## $ WorkPayCare                   <chr> "Strongly disagree", NA, "Disagree...
## $ KinshipDevelopers             <chr> "Agree", NA, "Somewhat agree", "St...
## $ ChallengeMyself               <chr> "Agree", NA, "Agree", "Strongly ag...
## $ CompetePeers                  <chr> "Disagree", NA, "Disagree", "Somew...
## $ ChangeWorld                   <chr> "Agree", NA, "Agree", "Agree", NA,...
## $ JobSeekingStatus              <chr> "I'm not actively looking, but I a...
## $ HoursPerWeek                  <int> 0, NA, NA, 5, NA, 0, 1, 1, 2, 1, N...
## $ LastNewJob                    <chr> "Not applicable/ never", NA, NA, "...
## $ AssessJobIndustry             <chr> "Very important", NA, NA, "Somewha...
## $ AssessJobRole                 <chr> "Very important", NA, NA, "Somewha...
## $ AssessJobExp                  <chr> "Important", NA, NA, "Somewhat imp...
## $ AssessJobDept                 <chr> "Very important", NA, NA, "Importa...
## $ AssessJobTech                 <chr> "Very important", NA, NA, "Importa...
## $ AssessJobProjects             <chr> "Very important", NA, NA, "Very im...
## $ AssessJobCompensation         <chr> "Important", NA, NA, "Important", ...
## $ AssessJobOffice               <chr> "Very important", NA, NA, "Very im...
## $ AssessJobCommute              <chr> "Very important", NA, NA, "Importa...
## $ AssessJobRemote               <chr> "Very important", NA, NA, "Somewha...
## $ AssessJobLeaders              <chr> "Very important", NA, NA, "Not ver...
## $ AssessJobProfDevel            <chr> "Very important", NA, NA, "Very im...
```

```
## $ AssessJobDiversity               <chr> "Somewhat important", NA, NA, "Imp...
## $ AssessJobProduct                 <chr> "Not very important", NA, NA, "Ver...
## $ AssessJobFinances                <chr> "Somewhat important", NA, NA, "Ver...
## $ ImportantBenefits                <chr> "Stock options; Vacation/days off;...
## $ ClickyKeys                       <chr> "Yes", "No", "Yes", "Yes", NA, "Ye...
## $ JobProfile                       <chr> "Other", "Other", NA, "LinkedIn; O...
## $ ResumePrompted                   <chr> NA, NA, NA, NA, NA, NA, NA, "A rec...
## $ LearnedHiring                    <chr> NA, "Some other way", NA, "A frien...
## $ ImportantHiringAlgorithms        <chr> "Important", "Important", NA, "Som...
## $ ImportantHiringTechExp           <chr> "Important", "Important", NA, "Som...
## $ ImportantHiringCommunication     <chr> "Important", "Important", NA, "Ver...
## $ ImportantHiringOpenSource        <chr> "Somewhat important", "Important",...
## $ ImportantHiringPMExp             <chr> "Important", "Somewhat important",...
## $ ImportantHiringCompanies         <chr> "Not very important", "Somewhat im...
## $ ImportantHiringTitles            <chr> "Not very important", "Not very im...
## $ ImportantHiringEducation         <chr> "Not at all important", "Somewhat ...
## $ ImportantHiringRep               <chr> "Somewhat important", "Not very im...
## $ ImportantHiringGettingThingsDone <chr> "Very important", "Very important"...
## $ Currency                         <chr> NA, "British pounds sterling (Â£)"...
## $ Overpaid                         <chr> NA, NA, "Neither underpaid nor ove...
## $ TabsSpaces                       <chr> "Tabs", "Spaces", "Spaces", "Space...
## $ EducationImportant               <chr> NA, NA, "Not very important", NA, ...
## $ EducationTypes                   <chr> "Online course; Open source contri...
## $ SelfTaughtTypes                  <chr> NA, "Official documentation; Stack...
## $ TimeAfterBootcamp                <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ CousinEducation                  <chr> NA, NA, NA, NA, NA, NA, NA, "Get a...
## $ WorkStart                        <chr> "6:00 AM", "10:00 AM", "9:00 AM", ...
## $ HaveWorkedLanguage               <chr> "Swift", "JavaScript; Python; Ruby...
## $ WantWorkLanguage                 <chr> "Swift", "Java; Python; Ruby; SQL"...
## $ HaveWorkedFramework              <chr> NA, ".NET Core", NA, "React", NA, ...
## $ WantWorkFramework                <chr> NA, ".NET Core", NA, "Hadoop; Node...
## $ HaveWorkedDatabase               <chr> NA, "MySQL; SQLite", "MySQL", "Mon...
## $ WantWorkDatabase                 <chr> NA, "MySQL; SQLite", NA, "MongoDB;...
## $ HaveWorkedPlatform               <chr> "iOS", "Amazon Web Services (AWS)"...
## $ WantWorkPlatform                 <chr> "iOS", "Linux Desktop; Raspberry P...
## $ IDE                              <chr> "Atom; Xcode", "Atom; Notepad++; V...
## $ AuditoryEnvironment              <chr> "Turn on some music", "Put on some...
## $ Methodology                      <chr> NA, NA, "Agile; Lean; Scrum; Extre...
## $ VersionControl                   <chr> NA, "Git", "Mercurial", "Git", NA,...
## $ CheckInCode                      <chr> NA, "Multiple times a day", "Multi...
## $ ShipIt                           <chr> NA, "Agree", "Agree", "Somewhat ag...
## $ OtherPeoplesCode                 <chr> NA, "Disagree", "Disagree", "Agree...
## $ ProjectManagement                <chr> NA, "Strongly disagree", "Disagree...
## $ EnjoyDebugging                   <chr> NA, "Agree", "Agree", "Somewhat ag...
## $ InTheZone                        <chr> NA, "Somewhat agree", "Agree", "St...
## $ DifficultCommunication           <chr> NA, "Disagree", "Disagree", "Disag...
## $ CollaborateRemote                <chr> NA, "Strongly disagree", "Somewhat...
## $ MetricAssess                     <chr> NA, "Customer satisfaction; On tim...
## $ EquipmentSatisfiedMonitors       <chr> "Somewhat satisfied", "Not very sa...
## $ EquipmentSatisfiedCPU            <chr> "Not very satisfied", "Satisfied",...
## $ EquipmentSatisfiedRAM            <chr> "Not at all satisfied", "Satisfied...
## $ EquipmentSatisfiedStorage        <chr> "Very satisfied", "Satisfied", "Sa...
## $ EquipmentSatisfiedRW             <chr> "Satisfied", "Somewhat satisfied",...
## $ InfluenceInternet                <chr> "Not very satisfied", "Satisfied",...
```

```
## $ InfluenceWorkstation            <chr> NA, "No influence at all", "A lot ...
## $ InfluenceHardware               <chr> NA, "No influence at all", "Some i...
## $ InfluenceServers                <chr> NA, "No influence at all", "Some i...
## $ InfluenceTechStack              <chr> NA, "No influence at all", "Some i...
## $ InfluenceDeptTech               <chr> NA, "No influence at all", "A lot ...
## $ InfluenceVizTools               <chr> NA, "No influence at all", "Some i...
## $ InfluenceDatabase               <chr> NA, "No influence at all", "Some i...
## $ InfluenceCloud                  <chr> NA, "No influence at all", "Some i...
## $ InfluenceConsultants            <chr> NA, "No influence at all", "Some i...
## $ InfluenceRecruitment            <chr> NA, "No influence at all", "Some i...
## $ InfluenceCommunication          <chr> NA, "No influence at all", "Some i...
## $ StackOverflowDescribes          <chr> "I have created a CV or Developer ...
## $ StackOverflowSatisfaction       <int> 9, 8, 8, 10, NA, 6, 8, 7, 8, 9, 10...
## $ StackOverflowDevices            <chr> "Desktop; iOS app", "Desktop; iOS ...
## $ StackOverflowFoundAnswer        <chr> "At least once each week", "Severa...
## $ StackOverflowCopiedCode         <chr> "Haven't done at all", "Several ti...
## $ StackOverflowJobListing         <chr> "Once or twice", "Once or twice", ...
## $ StackOverflowCompanyPage        <chr> "Haven't done at all", "Once or tw...
## $ StackOverflowJobSearch          <chr> "Haven't done at all", "Once or tw...
## $ StackOverflowNewQuestion        <chr> "Several times", "Haven't done at ...
## $ StackOverflowAnswer             <chr> "Several times", "Several times", ...
## $ StackOverflowMetaChat           <chr> "Once or twice", "At least once ea...
## $ StackOverflowAdsRelevant        <chr> "Somewhat agree", "Disagree", "Dis...
## $ StackOverflowAdsDistracting     <chr> "Strongly disagree", "Strongly dis...
## $ StackOverflowModeration         <chr> "Strongly disagree", "Strongly dis...
## $ StackOverflowCommunity          <chr> "Strongly agree", "Strongly agree"...
## $ StackOverflowHelpful            <chr> "Agree", "Agree", "Agree", "Strong...
## $ StackOverflowBetter             <chr> "Strongly agree", "Strongly agree"...
## $ StackOverflowWhatDo             <chr> "Strongly agree", "Strongly agree"...
## $ StackOverflowMakeMoney          <chr> "Strongly disagree", "Strongly dis...
## $ Gender                          <chr> "Male", "Male", "Male", "Male", NA...
## $ HighestEducationParents         <chr> "High school", "A master's degree"...
## $ Race                            <chr> "White or of European descent", "W...
## $ SurveyLong                      <chr> "Strongly disagree", "Somewhat agr...
## $ QuestionsInteresting            <chr> "Strongly agree", "Somewhat agree"...
## $ QuestionsConfusing              <chr> "Disagree", "Disagree", "Disagree"...
## $ InterestedAnswers               <chr> "Strongly agree", "Strongly agree"...
## $ Salary                          <dbl> NA, NA, 113750, NA, NA, NA, NA, NA...
## $ ExpectedSalary                  <dbl> NA, 37500, NA, NA, NA, NA, NA, NA,...
```

There is too many columns in this data set so I'm going to slim it down to 15 that seem more important than other columns

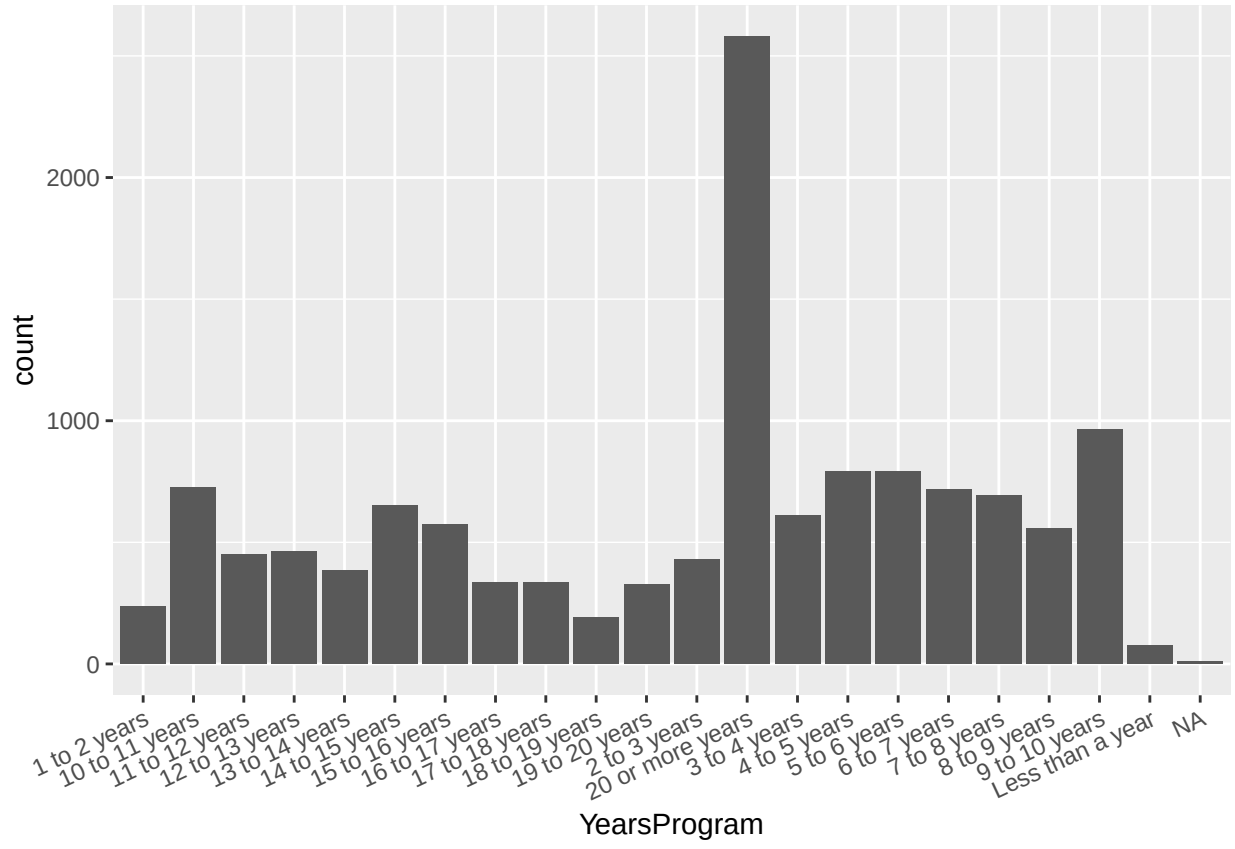
```
df <- data %>% drop_na(Salary) %>%
  select(Gender,Race,Salary,CompanySize,Country,YearsProgram,CareerSatisfaction,JobSatisfaction,HoursPe
         LearningNewTech)
glimpse(df)
```

```
## Rows: 12,891
## Columns: 15
## $ Gender             <chr> "Male", "Male", "Male", NA, "Male", "Male", "Mal...
## $ Race               <chr> "White or of European descent", "White or of Eur...
## $ Salary             <dbl> 113750.000, 100000.000, 130000.000, 82500.000, 1...
```

```
## $ CompanySize         <chr> "10,000 or more employees", "5,000 to 9,999 empl...
## $ Country             <chr> "United Kingdom", "United Kingdom", "United Stat...
## $ YearsProgram        <chr> "20 or more years", "20 or more years", "20 or m...
## $ CareerSatisfaction <int> 8, 8, 9, 5, 8, 7, 10, 7, NA, 6, 9, 10, 5, 8, 8, ...
## $ JobSatisfaction     <int> 9, 8, 8, 3, 9, 7, 8, 9, NA, 5, 9, 6, 5, 5, 8, 8,...
## $ HoursPerWeek        <int> NA, NA, NA, NA, NA, 0, 1, NA, 1, 4, NA, 2, NA, N...
## $ FormalEducation     <chr> "Bachelor's degree", "Professional degree", "Bac...
## $ EducationTypes      <chr> "Self-taught; Coding competition; Hackathon; Ope...
## $ ProblemSolving      <chr> "Strongly agree", "Strongly agree", "Strongly ag...
## $ JobSecurity         <chr> "Agree", "Somewhat agree", "Agree", "Strongly ag...
## $ VersionControl      <chr> "Mercurial", "Git", NA, NA, "Git", NA, "Team Fou...
## $ LearningNewTech     <chr> "Strongly agree", "Agree", "Strongly agree", "St...
```

Taking a look at distribution of years of programing

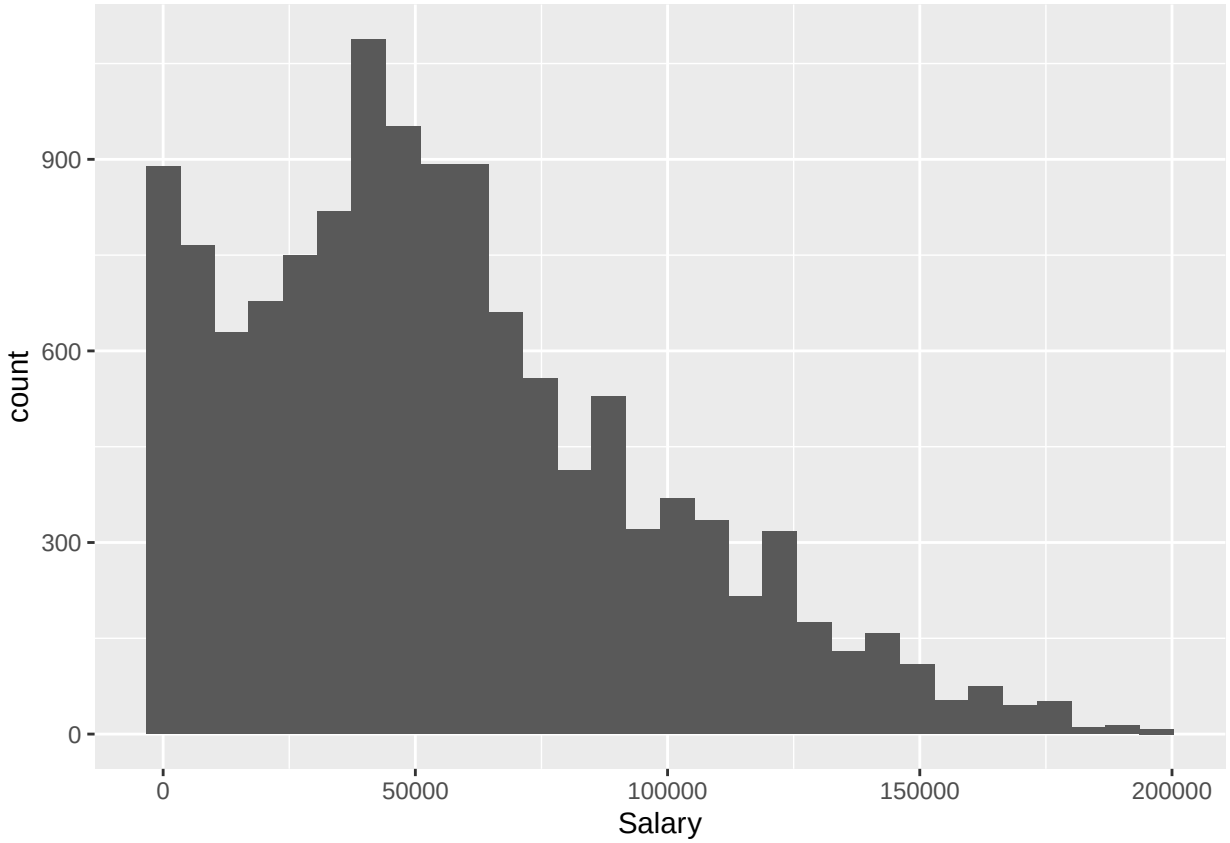
```
ggplot(df,
       aes(YearsProgram)
       ) + geom_bar()+ scale_x_discrete(guide = guide_axis(angle = 25))
```



looks like a vast majority of respondents have over 20 years of experience. Now I would like to see the distribution of salaries in the data set

```
ggplot(df,
       aes(Salary)
       ) + geom_histogram()
```

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



Appears I may have a lot of people who entered 0 for their salary. These could be students or people choosing to abstain from answering. I would like to see distributions of different demographic data.

```
df %>% group_by(Gender) %>% summarise(count = n()) %>% arrange(-count)
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 20 x 2
##    Gender                                                  count
##    <chr>                                                   <int>
##  1 Male                                                    10666
##  2 <NA>                                                     1182
##  3 Female                                                    819
##  4 Other                                                      49
##  5 Male; Other                                                44
##  6 Gender non-conforming                                      43
##  7 Male; Gender non-conforming                                21
##  8 Female; Transgender                                        19
##  9 Female; Gender non-conforming                              13
## 10 Transgender                                                 8
## 11 Male; Female; Transgender; Gender non-conforming; Other     6
## 12 Male; Female                                                5
## 13 Male; Transgender                                           4
## 14 Transgender; Gender non-conforming                          4
```

```
## 15 Female; Transgender; Gender non-conforming                  3
## 16 Female; Transgender; Other                                  1
## 17 Male; Female; Other                                         1
## 18 Male; Female; Transgender                                   1
## 19 Male; Gender non-conforming; Other                          1
## 20 Male; Transgender; Other                                    1
```

Looks like there is 20 different gender values but the data set is disproportionately male. These statistics will be biased towards men after reviewing this factor and should be considered when evaluating results.

```
as.data.frame(table(unlist(str_split(df$EducationTypes,'; '))))
```

```
##                        Var1 Freq
## 1                  Bootcamp  776
## 2        Coding competition 2029
## 3                 Hackathon 2450
## 4    Industry certification 1420
## 5       On-the-job training 4297
## 6             Online course 3954
## 7 Open source contributions 3576
## 8  Part-time/evening course 1192
## 9               Self-taught 8194
```

Wanting to understand how the respondents were educated. I need to extract values from strings since the column held multiple values. The data is more than double counted but gives an idea of how the respondents learned coding. A majority have self taught themselves at least once for job related skills.

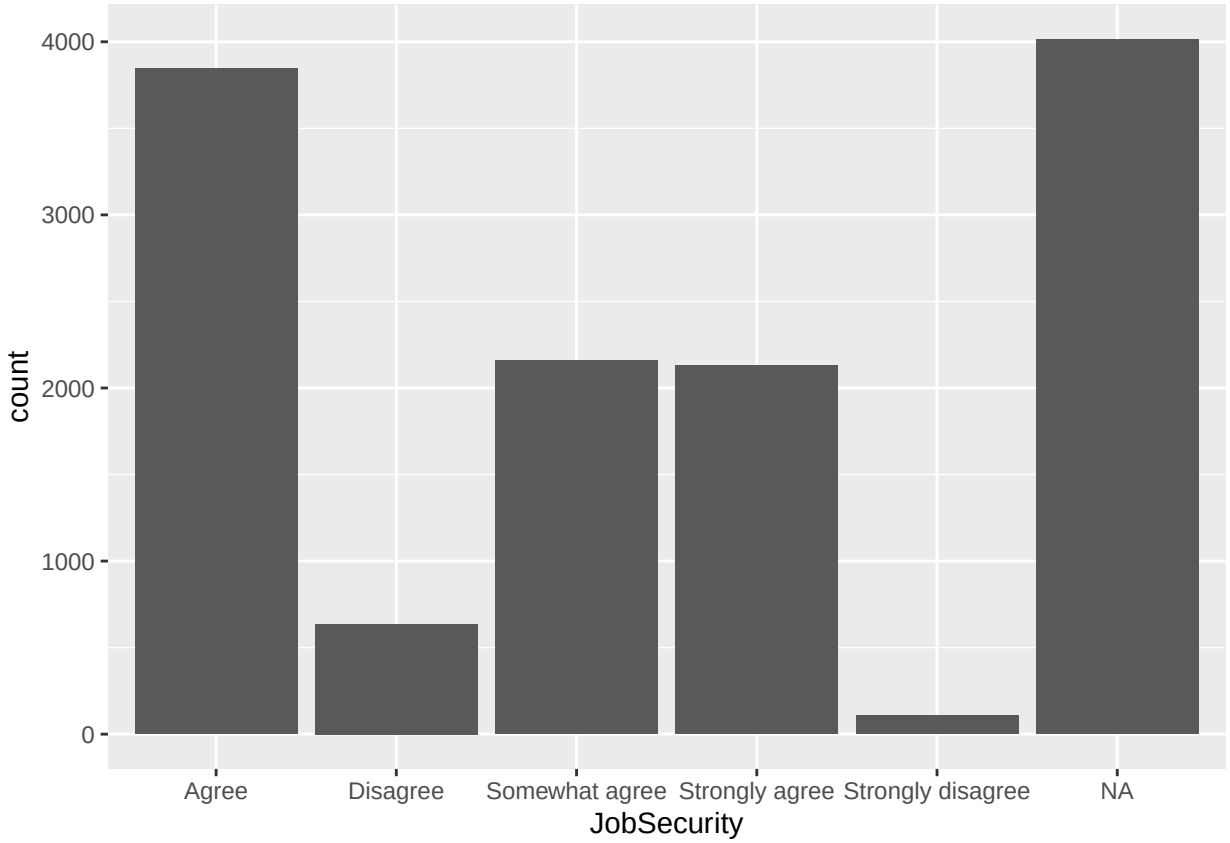
```
df %>% group_by(CompanySize) %>% summarise(count = n()) %>% arrange(-count)
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 11 x 2
##    CompanySize              count
##    <chr>                    <int>
##  1 20 to 99 employees        3065
##  2 100 to 499 employees      2610
##  3 10,000 or more employees  1823
##  4 10 to 19 employees        1337
##  5 1,000 to 4,999 employees  1288
##  6 Fewer than 10 employees   1184
##  7 500 to 999 employees       851
##  8 5,000 to 9,999 employees   520
##  9 I don't know               161
## 10 I prefer not to answer      41
## 11 <NA>                        11
```

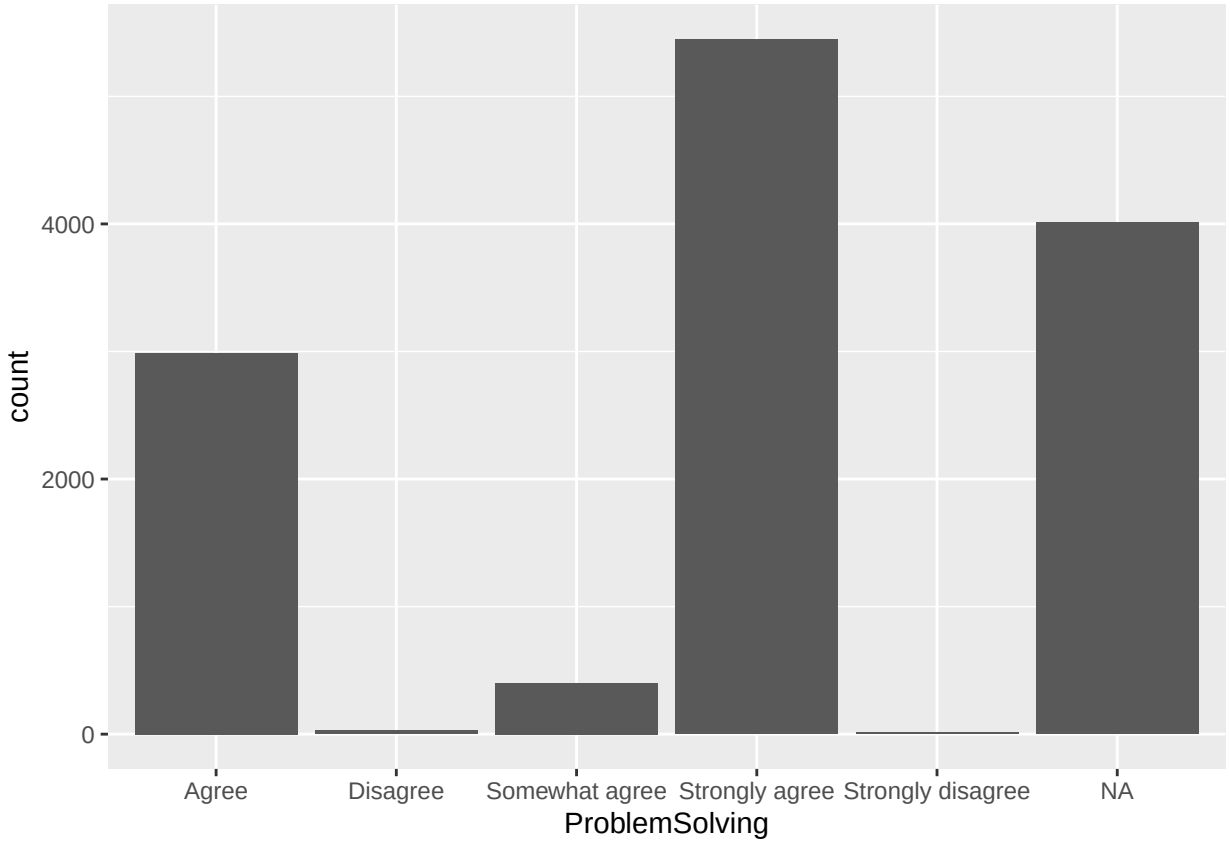
Appears we have a good distribution of company sizes ranging from less than 10 to over 10,000

```
ggplot(df,
       aes(JobSecurity)
       ) + geom_bar()
```



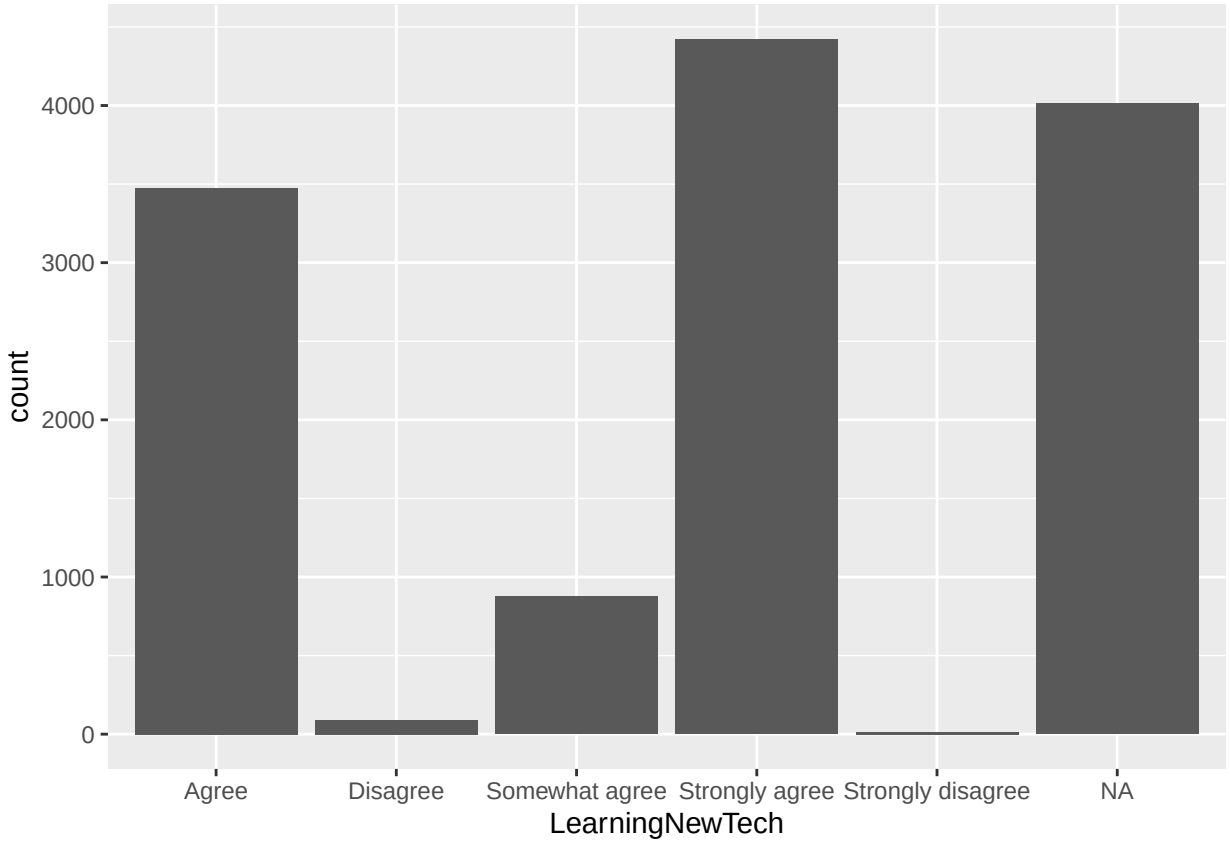
there is a lot of respondents who chose not to answer how important Job security is to their career. Of those that did respond, majority says they do agree job security is important.

```
ggplot(df,
       aes(ProblemSolving)
       ) + geom_bar()
```



Problem solving is extremely important for your career according to the respondents based on the graph above.

```
ggplot(df,
       aes(LearningNewTech)
       ) + geom_bar()
```



Learning new tech has very similar distribution to Problem solving

```
df %>% group_by(FormalEducation) %>% summarise(count = n()) %>% arrange(-count)
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 9 x 2
##   FormalEducation                                                count
##   <chr>                                                          <int>
## 1 Bachelor's degree                                               6407
## 2 Master's degree                                                 3077
## 3 Some college/university study without earning a bachelor's degree  2050
## 4 Secondary school                                                 761
## 5 Doctoral degree                                                  293
## 6 Professional degree                                              143
## 7 I never completed any formal education                            60
## 8 Primary/elementary school                                         55
## 9 I prefer not to answer                                            45
```

Even though most of the respondents stated that they are self taught I'm seeing that almost all respondents actually have at least some college education. So this data set is heavily favored for college educated respondents.

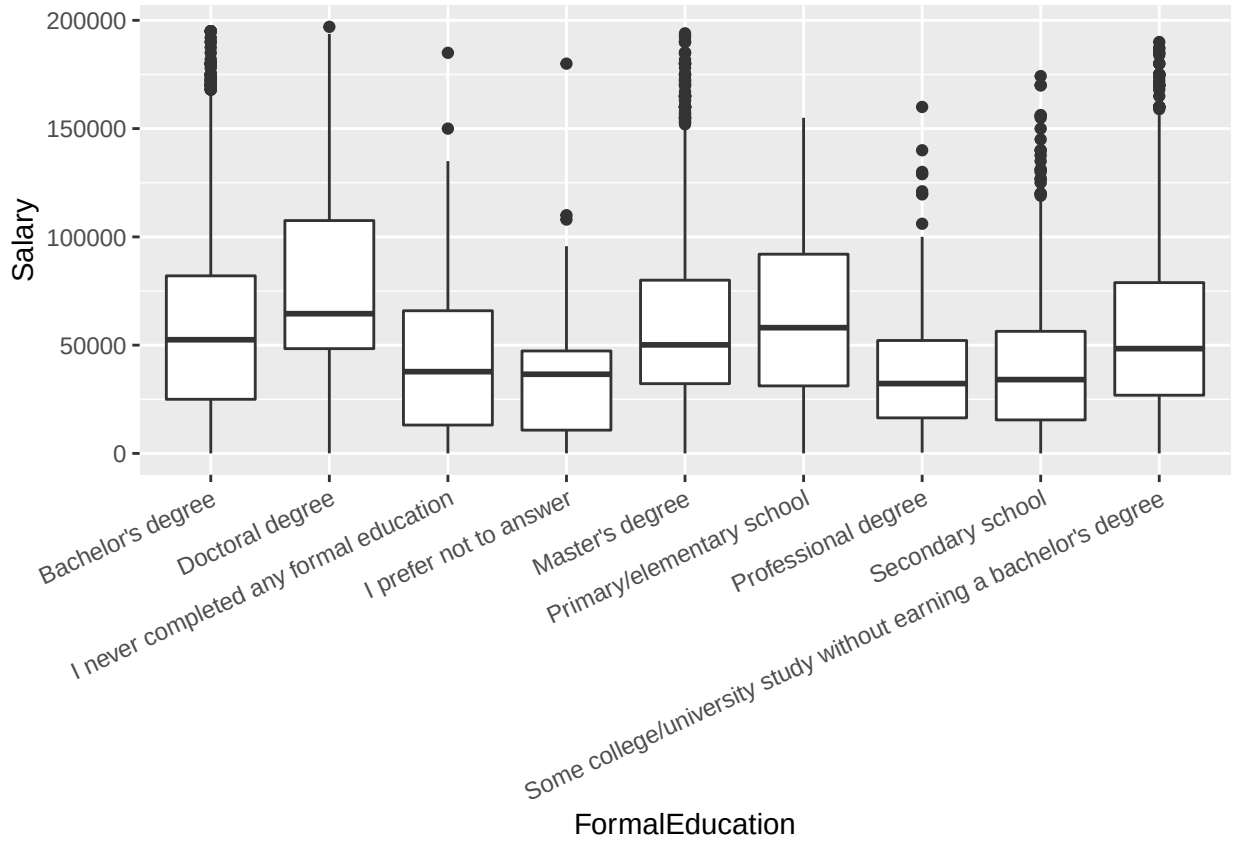
```
df %>% group_by(VersionControl) %>% summarise(count = n()) %>% arrange(-count)
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 11 x 2
##    VersionControl                               count
##    <chr>                                        <int>
##  1 Git                                           6410
##  2 <NA>                                          3984
##  3 Subversion                                     906
##  4 Team Foundation Server                         824
##  5 I use some other system                        300
##  6 Mercurial                                      194
##  7 I don't use version control                     80
##  8 Copying and pasting files to network shares     58
##  9 Zip file back-ups                               56
## 10 Rational ClearCase                              40
## 11 Visual Source Safe                              39
```

Majority of respondents use Github for version control which is not surprising.

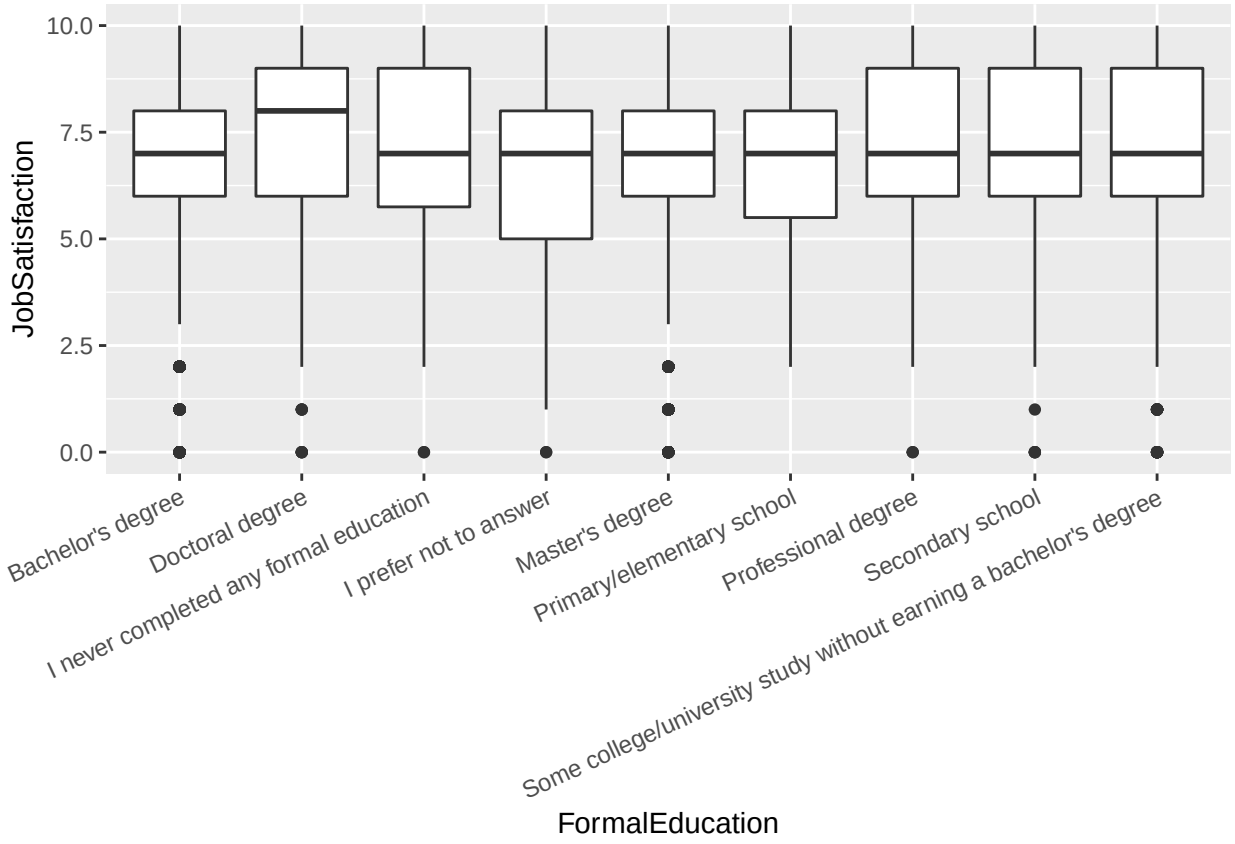
```
ggplot(df,
       aes(FormalEducation, Salary)
       ) + geom_boxplot() + scale_x_discrete(guide = guide_axis(angle = 25))
```



This one surprised me. It looks like a Masters Degree does not have higher salary ranges than bachelors degrees. Doctoral degrees do seem to have an impact however.

```
ggplot(df,
       aes(FormalEducation, JobSatisfaction)
       ) + geom_boxplot() + scale_x_discrete(guide = guide_axis(angle = 25))
```

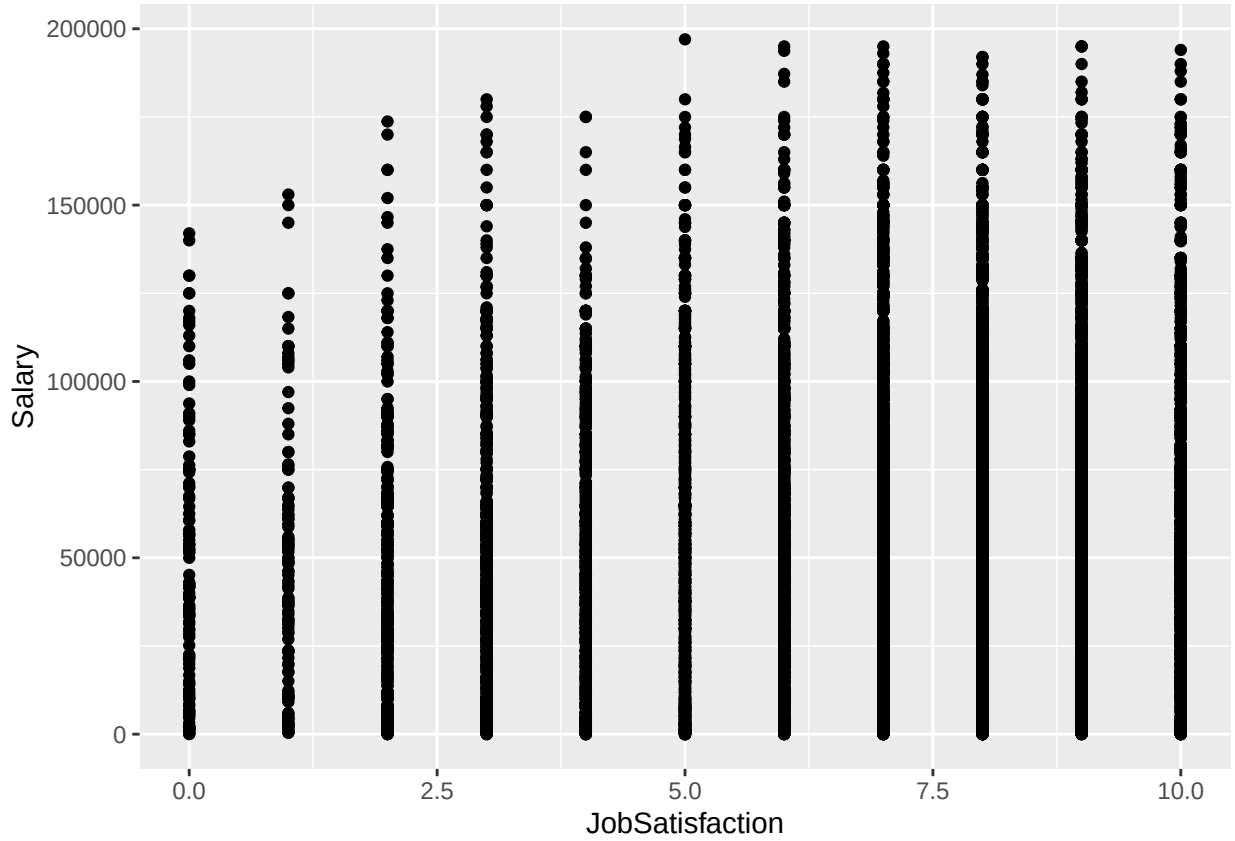
```
## Warning: Removed 39 rows containing non-finite values (stat_boxplot).
```



Seems college graduates have the lowest range of job satisfaction amongst respondents.

```
ggplot(df,
       aes(JobSatisfaction, Salary)
       ) + geom_point()
```

```
## Warning: Removed 39 rows containing missing values (geom_point).
```



Salary seems to be positively correlated with job satisfaction

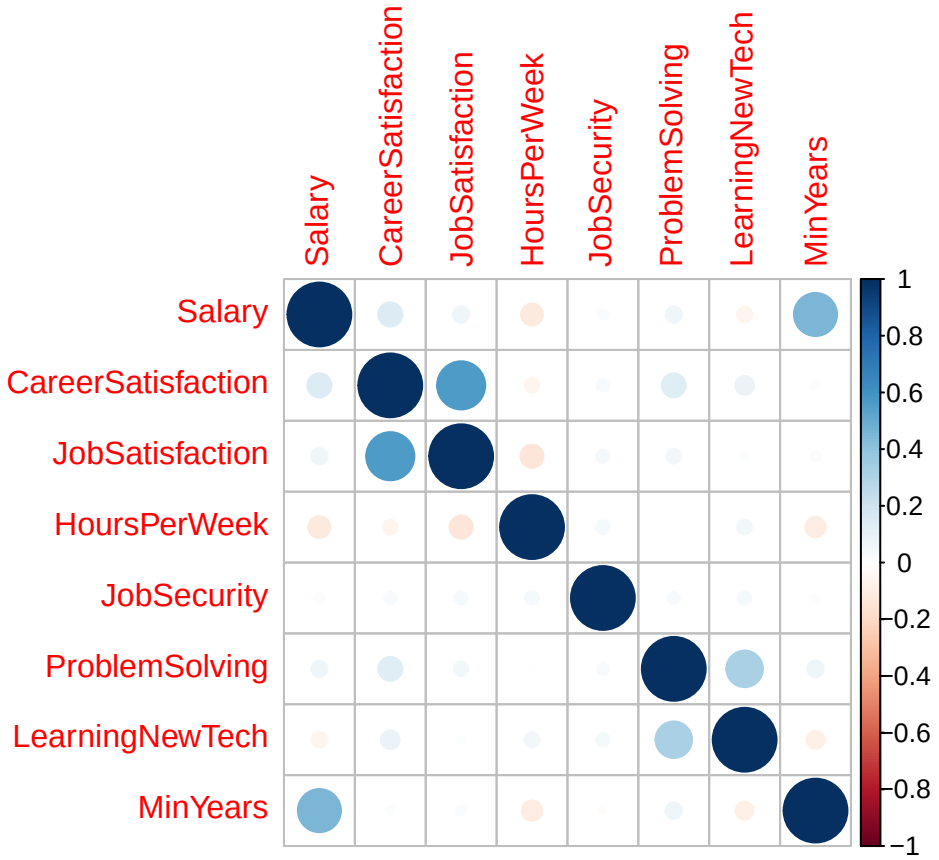
I'm going to change some of the variables to numerical values to do some better analysis. First I'll change in range of agree to disagree variables to a 1-5 scale. Then I'm going to change years of programing experiencing to a min year numerical value.

```
df <- df %>% mutate(JobSecurity = case_when(df$JobSecurity == 'Strongly disagree' ~ 1,
                                df$JobSecurity == 'Disagree' ~ 2,
                                df$JobSecurity == 'Somewhat agree' ~ 3,
                                df$JobSecurity == 'Agree' ~ 4,
                                df$JobSecurity == 'Strongly agree' ~ 5),
                    ProblemSolving = case_when(df$ProblemSolving == 'Strongly disagree' ~ 1,
                                df$ProblemSolving == 'Disagree' ~ 2,
                                df$ProblemSolving == 'Somewhat agree' ~ 3,
                                df$ProblemSolving == 'Agree' ~ 4,
                                df$ProblemSolving == 'Strongly agree' ~ 5),
                    LearningNewTech = case_when(df$LearningNewTech == 'Strongly disagree' ~ 1,
                                df$LearningNewTech == 'Disagree' ~ 2,
                                df$LearningNewTech == 'Somewhat agree' ~ 3,
                                df$LearningNewTech == 'Agree' ~ 4,
                                df$LearningNewTech == 'Strongly agree' ~ 5))

df$MinYears <- as.integer(str_extract(df$YearsProgram,'(\\d)+'))

df <- df %>% filter(Salary > 0)

glimpse(df)
```

```
## Rows: 12,885
## Columns: 16
## $ Gender              <chr> "Male", "Male", "Male", NA, "Male", "Male", "Mal...
## $ Race                <chr> "White or of European descent", "White or of Eur...
## $ Salary              <dbl> 113750.000, 100000.000, 130000.000, 82500.000, 1...
## $ CompanySize         <chr> "10,000 or more employees", "5,000 to 9,999 empl...
## $ Country             <chr> "United Kingdom", "United Kingdom", "United Stat...
## $ YearsProgram        <chr> "20 or more years", "20 or more years", "20 or m...
## $ CareerSatisfaction <int> 8, 8, 9, 5, 8, 7, 10, 7, NA, 6, 9, 10, 5, 8, 8, ...
## $ JobSatisfaction     <int> 9, 8, 8, 3, 9, 7, 8, 9, NA, 5, 9, 6, 5, 5, 8, 8,...
## $ HoursPerWeek        <int> NA, NA, NA, NA, NA, 0, 1, NA, 1, 4, NA, 2, NA, N...
## $ FormalEducation     <chr> "Bachelor's degree", "Professional degree", "Bac...
## $ EducationTypes      <chr> "Self-taught; Coding competition; Hackathon; Ope...
## $ ProblemSolving      <dbl> 5, 5, 5, 5, NA, 4, NA, 5, 5, 5, NA, NA, 4, 4, 5,...
## $ JobSecurity         <dbl> 4, 3, 4, 5, NA, 4, NA, 4, 4, 5, NA, NA, 5, 3, 3,...
## $ VersionControl      <chr> "Mercurial", "Git", NA, NA, "Git", NA, "Team Fou...
## $ LearningNewTech     <dbl> 5, 4, 5, 5, NA, 5, NA, 5, 4, 5, NA, NA, 4, 4, 4,...
## $ MinYears            <int> 20, 20, 20, 2, 10, 20, 7, 20, 20, 4, 8, 11, 3, 1...
```

Now that I have a number of numerical columns I can create a corplot to see any correlations.

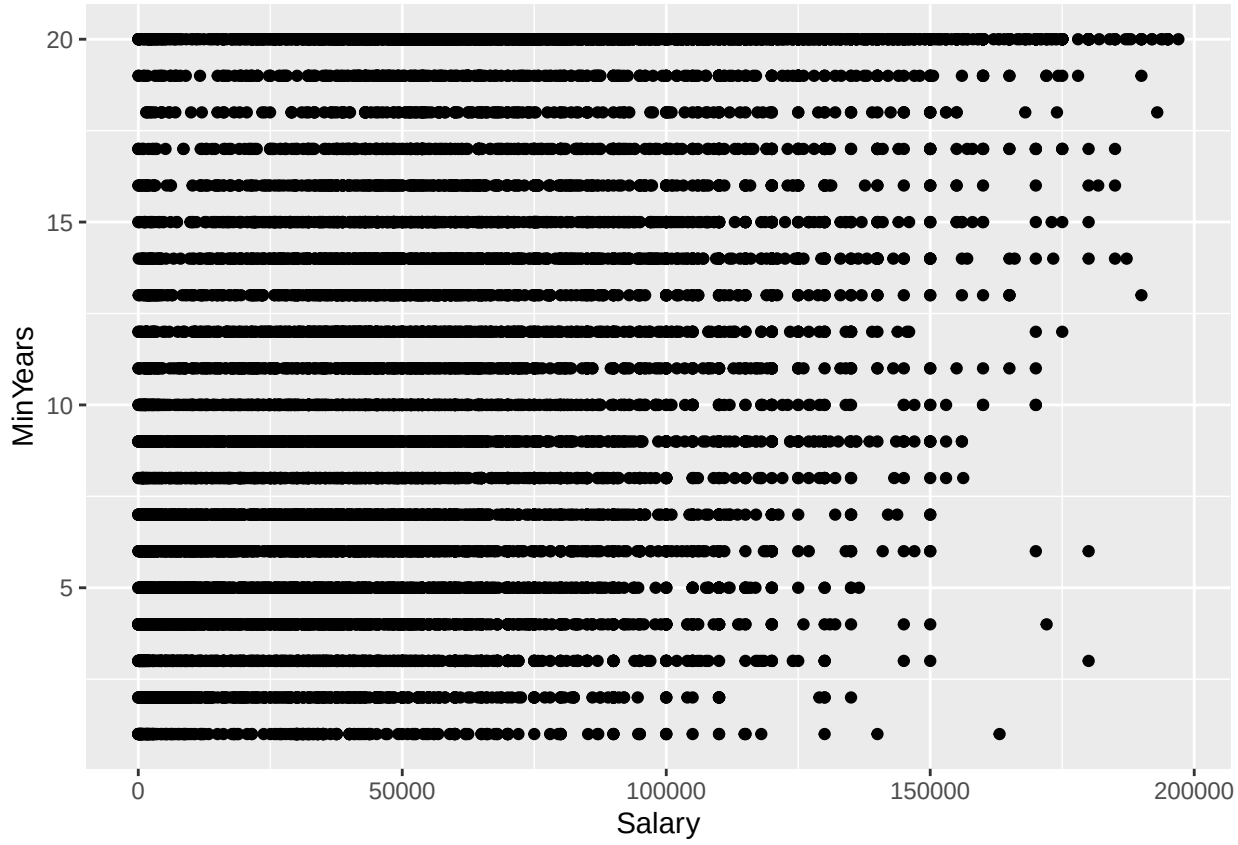
```
core <- df %>% select(Salary,CareerSatisfaction,JobSatisfaction,HoursPerWeek,JobSecurity,ProblemSolving

cor(core) %>%
  corrplot(.)
```



It looks like there is strong correlation between Job and Career satisfaction. A strong correlation between Years of experience and salary. A moderate correlation between Learning New technologies and problem solving.

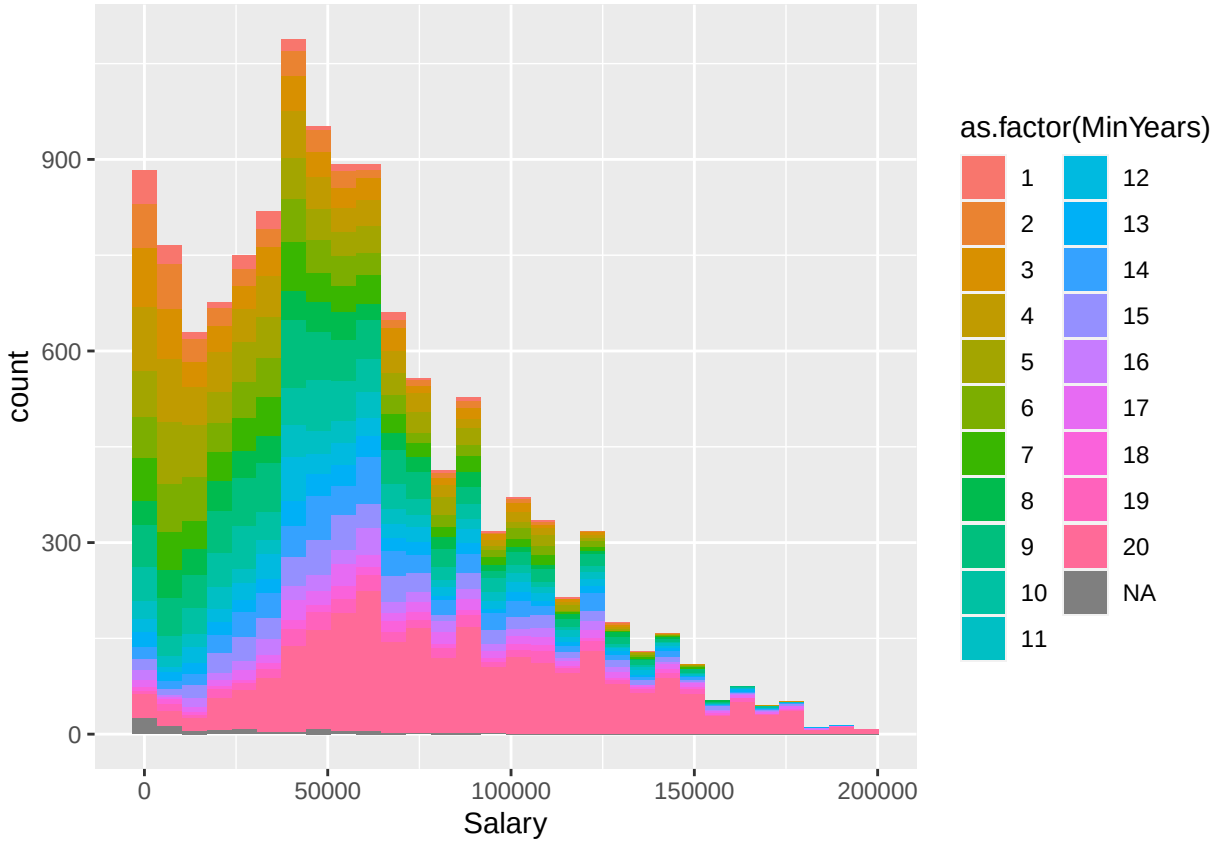
```
ggplot(df,
       aes(Salary,MinYears)
       ) +geom_point()
```

```
## Warning: Removed 86 rows containing missing values (geom_point).
```




```
ggplot(df,
       aes(Salary, fill=as.factor(MinYears))
       ) + geom_histogram()
```

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



When looking at salary based on years of experience I can see that almost all respondents making over 150K have at least 20 years of experience.