Data Exercise for NTLAKIS

2024-03-02

Data Cleaning and Preparations

```
glimpse(df)
```

```
## Rows: 3,737
## Columns: 24
## $ `Employee ID`         <dbl> 133315, 133551, 134280, 137160, 141662, 144151, ~
## $ Gender                <chr> "F", "M", "M", "M", "M", "M", "M", "F", "F", "F"~
## $ `Race/ethnicity`      <chr> "Hispanic or Latino", "White", "Hispanic or Lati~
## $ `Job Profile`         <chr> "Vice President Regulatory Affairs", "Sr Legal C~
## $ `Job Group`           <chr> "1A", "1B", "1A", "1B", "1A", "1B", "1B", "1A", ~
## $ `Job Group Name`      <chr> "VPs and Above", "Directors", "VPs and Above", "~
## $ `EEO-1 Category`      <chr> "1.1 - Executive / Sr. Level Officials and Manag~
## $ `FLSA Status`         <chr> "EX", "EX", "EX", "EX", "EX", "EX", "EX", "EX", ~
## $ Grade                 <chr> "L3", "L1", "L2", "L1", "L3", "L1", "L1", "L3", ~
## $ `Job Family`          <chr> "EXC", "LEG", "EXC", "SLS", "EXC", "LEG", "BD", ~
## $ Department            <chr> "Regulatory", "Intellectual Property", "Corporat~
## $ `Annual Pay`          <dbl> 254269.6, 249378.6, 269254.7, 258560.1, 255393.8~
## $ FTE                   <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ~
## $ `Pay Type`            <chr> "Salary", "Salary", "Salary", "Salary", "Salary"~
## $ `Job Entry Date`      <dttm> 2017-07-17, 2002-10-17, 2018-02-14, 2010-06-16,~
## $ `Hire Date`           <dttm> 1985-12-29, 1992-04-26, 1999-01-17, 1991-05-17,~
## $ `Date of Birth`       <dttm> 1960-03-26, 1968-02-26, 1976-04-15, 1967-05-24,~
## $ `Location ID`         <chr> "NYC", "BOS", "NYC", "BOS", "BOS", "BOS", "BOS",~
## $ `Salary Plan`         <dbl> 1.2, 1.2, 1.2, 1.2, 1.2, 1.2, 1.2, 1.2, 1.2, 0.9~
## $ Education             <chr> "Not Indicated", "Doctorate Degree", "Master's D~
## $ `FY 2019 Rating`      <chr> "Meets Expectations", "Meets Expectations", "Top~
## $ `FY 2018 Performance` <dbl> 3, 3, 5, 5, 3, 4, 4, 4, 3, 3, 3, 3, 4, 3, 3, 4, ~
## $ `FY 2017 Performance` <dbl> 3, 4, NA, 3, 3, 3, NA, 3, 4, 3, 3, 4, 5, 3, 3, 3~
## $ `FY 2016 Performance` <dbl> 4, 3, NA, 5, 3, 3, NA, 4, 2, 3, 3, 3, 4, 3, 4, 3~
```

```
summary(df)
```

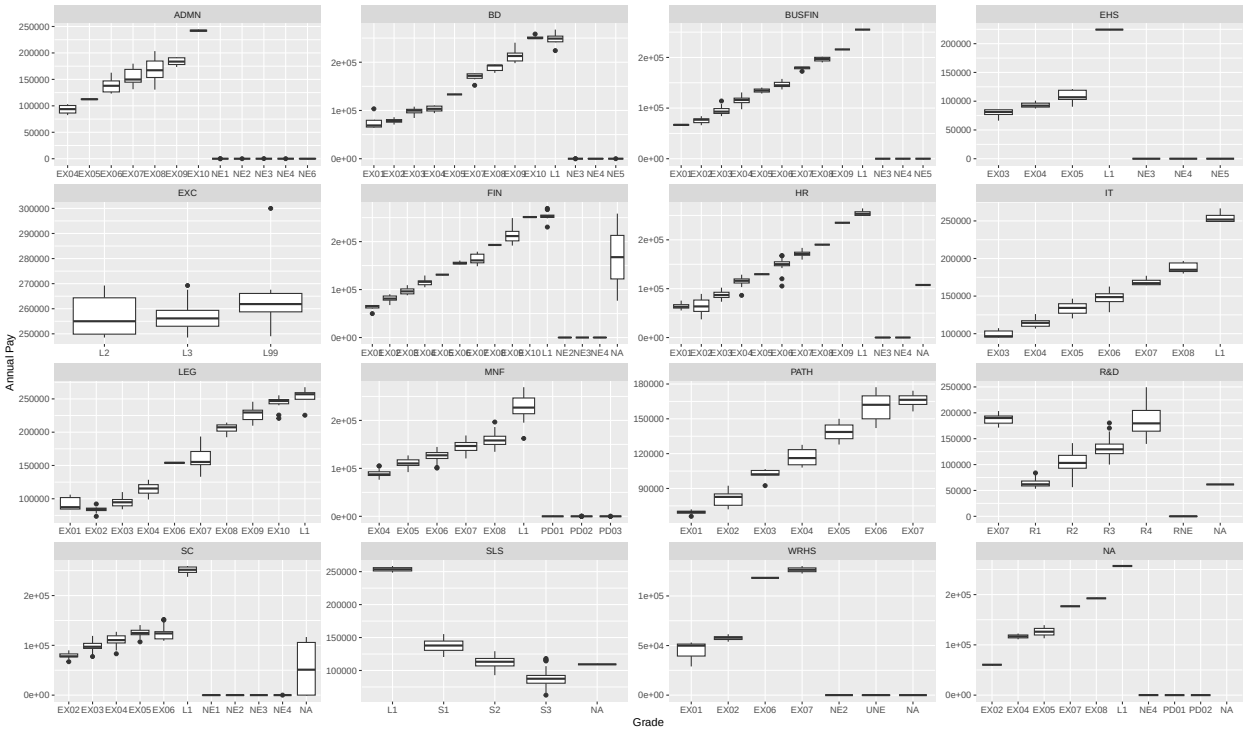
```
##   Employee ID        Gender          Race/ethnicity     Job Profile
##  Min.   :128551   Length:3737        Length:3737        Length:3737
##  1st Qu.:344829   Class :character   Class :character   Class :character
##  Median :551620   Mode  :character   Mode  :character   Mode  :character
##  Mean   :552474
##  3rd Qu.:766127
##  Max.   :972017
##
##   Job Group         Job Group Name     EEO-1 Category     FLSA Status
##  Length:3737        Length:3737        Length:3737        Length:3737
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
```

```
##
##
##
##     Grade             Job Family         Department          Annual Pay
##  Length:3737        Length:3737        Length:3737        Min.   :     2.10
##  Class :character   Class :character   Class :character   1st Qu.:    27.13
##  Mode  :character   Mode  :character   Mode  :character   Median : 92585.35
##                                                           Mean   : 85737.67
##                                                           3rd Qu.:131463.54
##                                                           Max.   :300000.00
##                                                           NA's   :8
##       FTE           Pay Type         Job Entry Date
##  Min.   :0.5000   Length:3737        Min.   :1986-06-16 00:00:00.00
##  1st Qu.:1.0000   Class :character   1st Qu.:2013-06-16 00:00:00.00
##  Median :1.0000   Mode  :character   Median :2016-05-17 00:00:00.00
##  Mean   :0.9965                      Mean   :2014-11-26 02:38:42.76
##  3rd Qu.:1.0000                      3rd Qu.:2017-09-01 00:00:00.00
##  Max.   :1.0000                      Max.   :2018-12-16 00:00:00.00
##  NA's   :8                           NA's   :8
##    Hire Date                     Date of Birth
##  Min.   :1968-04-25 00:00:00.00   Min.   :1900-01-01 00:00:00.0000
##  1st Qu.:1997-03-02 00:00:00.00   1st Qu.:1965-03-07 00:00:00.0000
##  Median :2008-03-17 00:00:00.00   Median :1970-05-16 00:00:00.0000
##  Mean   :2005-03-27 13:52:42.64   Mean   :1972-09-14 19:47:50.1528
##  3rd Qu.:2015-05-15 00:00:00.00   3rd Qu.:1979-06-24 00:00:00.0000
##  Max.   :2018-12-16 00:00:00.00   Max.   :1999-06-20 00:00:00.0000
##                                   NA's   :8
##  Location ID         Salary Plan    Education         FY 2019 Rating
##  Length:3737        Min.   :0.90   Length:3737        Length:3737
##  Class :character   1st Qu.:1.00   Class :character   Class :character
##  Mode  :character   Median :1.00   Mode  :character   Mode  :character
##                     Mean   :1.07
##                     3rd Qu.:1.20
##                     Max.   :1.20
##                     NA's   :8
##  FY 2018 Performance FY 2017 Performance FY 2016 Performance
##  Min.   :1.000       Min.   :1.000       Min.   :1.000
##  1st Qu.:3.000       1st Qu.:3.000       1st Qu.:3.000
##  Median :3.000       Median :3.000       Median :3.000
##  Mean   :3.282       Mean   :3.289       Mean   :3.328
##  3rd Qu.:4.000       3rd Qu.:4.000       3rd Qu.:4.000
##  Max.   :5.000       Max.   :5.000       Max.   :5.000
##  NA's   :1390        NA's   :1523        NA's   :1762
```

Investigating the unusual range observed in the Annual Salary variable, an examination for potential extreme outliers.

```
ggplot(df, aes(x = `Grade`, y = `Annual Pay`)) +
  geom_boxplot() +
  facet_wrap(.~`Job Family`, scales = "free", nrow = 4, ncol = 4)
```

```
## Warning: Removed 8 rows containing non-finite values (`stat_boxplot()`).
```



There are no discernible extreme outliers when examined with respect to both grade and job family levels.

Check if there is any formatting issue or discrepancies in the data.

```
column_types <- sapply(df, class)

# Print the result
print(column_types)
```

```
## $`Employee ID`
## [1] "numeric"
##
## $Gender
## [1] "character"
##
## $`Race/ethnicity`
## [1] "character"
##
## $`Job Profile`
## [1] "character"
##
## $`Job Group`
## [1] "character"
##
## $`Job Group Name`
## [1] "character"
##
## $`EEO-1 Category`
## [1] "character"
##
## $`FLSA Status`
```

```
## [1] "character"
##
## $Grade
## [1] "character"
##
## $`Job Family`
## [1] "character"
##
## $Department
## [1] "character"
##
## $`Annual Pay`
## [1] "numeric"
##
## $FTE
## [1] "numeric"
##
## $`Pay Type`
## [1] "character"
##
## $`Job Entry Date`
## [1] "POSIXct" "POSIXt"
##
## $`Hire Date`
## [1] "POSIXct" "POSIXt"
##
## $`Date of Birth`
## [1] "POSIXct" "POSIXt"
##
## $`Location ID`
## [1] "character"
##
## $`Salary Plan`
## [1] "numeric"
##
## $Education
## [1] "character"
##
## $`FY 2019 Rating`
## [1] "character"
##
## $`FY 2018 Performance`
## [1] "numeric"
##
## $`FY 2017 Performance`
## [1] "numeric"
##
## $`FY 2016 Performance`
## [1] "numeric"
```

Checking for Duplicates

```
# Want to check for duplicates based 'Employee ID'
duplicates <- df[duplicated('Employee ID'), ]
```

```
print(duplicates)
```

```
## # A tibble: 0 x 24
## # i 24 variables: Employee ID <dbl>, Gender <chr>, Race/ethnicity <chr>,
## #   Job Profile <chr>, Job Group <chr>, Job Group Name <chr>,
## #   EEO-1 Category <chr>, FLSA Status <chr>, Grade <chr>, Job Family <chr>,
## #   Department <chr>, Annual Pay <dbl>, FTE <dbl>, Pay Type <chr>,
## #   Job Entry Date <dttm>, Hire Date <dttm>, Date of Birth <dttm>,
## #   Location ID <chr>, Salary Plan <dbl>, Education <chr>,
## #   FY 2019 Rating <chr>, FY 2018 Performance <dbl>, ...
```

Performing assessment of null identifiers with potential replacements and scrutinizing unconventional entries in certain columns:

Check for unusual Gender identifiers.

```
unique(df$`Gender`)
```

```
## [1] "F"             "M"             "Not Disclosed" "Not Specified"
## [5] NA
```

We will replace the NA values with "Not Disclosed"

```
df <- df %>%
  mutate(Gender = replace_na(Gender, "Not Disclosed"))
```

Check for unusual Race/ethnicity identifiers.

```
unique(df$`Race/ethnicity`)
```

```
## [1] "Hispanic or Latino"             "White"
## [3] "Asian"                          "Black or African American"
## [5] "Two or More Races"              NA
## [7] "Not Specified"                  "American Indian/Alaskan Native"
```

Replace null values with Not Specified.

```
df <- df %>%
  mutate(`Race/ethnicity` = replace_na(`Race/ethnicity`, "Not Specified"))
```

NA's are present in following important columns: "Annual Pay","FTE", "Job Entry Date", "Date Of Birth", "Salary Plan". It is noteworthy that these columns exhibit an identical count of null values. Let us proceed to examine and analyze these null values.

```
# Filter out the columns of interest.
df %>%
  filter(is.na(`Annual Pay`) | is.na(FTE) |
           is.na(`Job Entry Date`) | is.na(`Date of Birth`) |
           is.na(`Salary Plan`))
```

```
## # A tibble: 8 x 24
##   `Employee ID` Gender `Race/ethnicity` `Job Profile` `Job Group`
##           <dbl> <chr>  <chr>            <chr>         <chr>
## 1        165960 M      Asian            Unassigned    <NA>
## 2        401217 F      White            Unassigned    <NA>
## 3        426147 F      White            Unassigned    <NA>
## 4        473192 F      White            Unassigned    <NA>
```

```
## 5        540124 F      White            Unassigned    <NA>
## 6        615117 F      White            Unassigned    <NA>
## 7        716371 M      White            Unassigned    <NA>
## 8        875969 M      White            Unassigned    <NA>
## # i 19 more variables: `Job Group Name` <chr>, `EEO-1 Category` <chr>,
## #   `FLSA Status` <chr>, Grade <chr>, `Job Family` <chr>, Department <chr>,
## #   `Annual Pay` <dbl>, FTE <dbl>, `Pay Type` <chr>, `Job Entry Date` <dttm>,
## #   `Hire Date` <dttm>, `Date of Birth` <dttm>, `Location ID` <chr>,
## #   `Salary Plan` <dbl>, Education <chr>, `FY 2019 Rating` <chr>,
## #   `FY 2018 Performance` <dbl>, `FY 2017 Performance` <dbl>,
## #   `FY 2016 Performance` <dbl>
```

```
df <- df %>%
  filter(!is.na(`Annual Pay`) & !is.na(FTE) &
           !is.na(`Job Entry Date`) & !is.na(`Date of Birth`) &
           !is.na(`Salary Plan`))
```

Dropping these rows are considered. The prevalence of null values across multiple columns suggests their limited utility, making their retention unnecessary for meaningful analysis.

Identify discrepancies in date entries with the following procedure:

```
# Filter out rows where Hire Date is greater than Job Entry Date
wrongdate <- df %>%
  filter(df$`Hire Date` > df$`Job Entry Date`)

print(wrongdate)
```

```
## # A tibble: 0 x 24
## # i 24 variables: Employee ID <dbl>, Gender <chr>, Race/ethnicity <chr>,
## #   Job Profile <chr>, Job Group <chr>, Job Group Name <chr>,
## #   EEO-1 Category <chr>, FLSA Status <chr>, Grade <chr>, Job Family <chr>,
## #   Department <chr>, Annual Pay <dbl>, FTE <dbl>, Pay Type <chr>,
## #   Job Entry Date <dttm>, Hire Date <dttm>, Date of Birth <dttm>,
## #   Location ID <chr>, Salary Plan <dbl>, Education <chr>,
## #   FY 2019 Rating <chr>, FY 2018 Performance <dbl>, ...
```

No anomalies detected in the date entries.

Let's now ensure comprehensive coverage by verifying the presence of any additional relevant null values that may have been overlooked.

```
null_counts <- colSums(is.na(df))

# Print the result
print(null_counts)
```

```
##         Employee ID              Gender      Race/ethnicity         Job Profile
##                   0                   0                   0                   0
##           Job Group      Job Group Name      EEO-1 Category         FLSA Status
##                   0                   0                   0                   0
##               Grade          Job Family          Department          Annual Pay
##                  14                  11                   0                   0
##                 FTE            Pay Type      Job Entry Date           Hire Date
##                   0                   0                   0                   0
```

```
##       Date of Birth         Location ID         Salary Plan           Education
##                   0                  10                   0                   0
##      FY 2019 Rating FY 2018 Performance FY 2017 Performance FY 2016 Performance
##                1382                1382                1515                1754
```

While null values are present, there is no immediate need for replacement or the removal of entire rows.

Finish Setting the Data.

Calculate Time spent in Company

```
lastday <- as.Date("2018-12-31")

df$`Time In Company(Years)` <- as.numeric(difftime(lastday, df$`Hire Date`, units = "days") /365)
#Take the difference between Last day and Hire date by days,then divide by 365 to get number of years
```

```
df %>%
  select(c('Employee ID', 'Time In Company(Years)'))
```

```
## # A tibble: 3,729 x 2
##    `Employee ID` `Time In Company(Years)`
##            <dbl>                    <dbl>
##  1        133315                     33.0
##  2        133551                     26.7
##  3        134280                     20.0
##  4        137160                     27.6
##  5        141662                     14.6
##  6        144151                     28.7
##  7        151024                     10.5
##  8        152320                     24.9
##  9        159737                     31.0
## 10        175634                     24.4
## # i 3,719 more rows
```

Make Dummy Integer

```
df$`Gender Dummy` <- ifelse(df$`Gender` == "M", 0, #If gender is M (male), then code it 0
                            ifelse(df$`Gender` == "F", 1, NA))
#If gender if F (female), then code it 1
```

```
df %>%
  select(c('Employee ID', 'Gender', 'Gender Dummy'))
```

```
## # A tibble: 3,729 x 3
##    `Employee ID` Gender `Gender Dummy`
##            <dbl> <chr>           <dbl>
##  1        133315 F                   1
##  2        133551 M                   0
##  3        134280 M                   0
##  4        137160 M                   0
##  5        141662 M                   0
##  6        144151 M                   0
##  7        151024 M                   0
##  8        152320 F                   1
```

```
##  9        159737 F                    1
## 10        175634 F                    1
## # i 3,719 more rows
```

Annualized base pay

```
df$`Annualized Base Pay` <- ifelse(df$`Pay Type` == "Hourly", 2080 * df$`Annual Pay`,
                                   #If Pay Type is hourly, multiply annual pay by 2080
                                ifelse(df$`Pay Type` == "Salary", (df$`Annual Pay` / df$FTE), NA))
#If Pay Type if salary, divide annual pay by FTE

df %>%
  select (c('Pay Type', 'Annualized Base Pay')) %>%
  slice_head(n=10)
```

```
## # A tibble: 10 x 2
##    `Pay Type` `Annualized Base Pay`
##    <chr>                      <dbl>
##  1 Salary                   254270.
##  2 Salary                   249379.
##  3 Salary                   269255.
##  4 Salary                   258560.
##  5 Salary                   255394.
##  6 Salary                   259343.
##  7 Salary                   267573.
##  8 Salary                   258358.
##  9 Salary                   249986.
## 10 Salary                   196878.
```

Generate a pivot table of the average Annualized base pay by gender

```
pivot_table <- df %>%
  group_by(`Gender Dummy`) %>%
  filter(!is.na(`Gender Dummy`)) %>% #omit non disclosed and not specified
  summarise(Average= mean(`Annualized Base Pay`)) %>% #averaged annual base pay
  pivot_wider(names_from = `Gender Dummy`, values_from = Average) %>%
  rename(Male = `0`, Female = `1`)

print(pivot_table)
```

```
## # A tibble: 1 x 2
##      Male Female
##     <dbl>  <dbl>
## 1 106828. 95889.
```

A t-test on Annualized base pay by sex:

```
ttest <- df %>%
  filter(`Grade` == "NE3" & `Job Family` == "ADMN" & !is.na(`Annualized Base Pay`))
```

```r
t.test(`Annualized Base Pay` ~ `Gender Dummy`, data = ttest)
```

```
##
##  Welch Two Sample t-test
##
## data:  Annualized Base Pay by Gender Dummy
## t = 0.97927, df = 4.0018, p-value = 0.3829
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -157042.2  328232.4
## sample estimates:
## mean in group 0 mean in group 1
##        150533.2         64938.1
```

Technical

In conducting a two-sample t-test, I have articulated two hypotheses to scrutinize the mean difference in annualized base salary between genders. The **null hypothesis** asserts that there is no significant difference between the mean of the annualized base salary between genders, while the **alternative hypothesis** that there is a significant difference between the mean of the annualized base salary between genders. The results indicate a t-score of 0.97927 and a corresponding p-value of 0.3829, suggesting *insufficient evidence to reject the null hypothesis* based on a significance level at 5%. The confidence interval, with a zero within its bounds, further fortifies the statistical argument in favor of the null hypothesis, indicating no notable divergence in the means of annualized base salary. As a result, we fail to reject the null hypothesis in this analysis.

Non Technical

In order to investigate whether there are any disparities in annual salaries between men and women, I conducted a statistical test. The core idea I explored was whether the *average annual base salary for men is essentially the same as that for women.* The results of this test revealed that the observed differences in average salaries between the two genders could happen approximately 38.29% of the time purely by random chance. Now, this moderate probability indicates that there **isn't strong enough evidence to confidently support the claim that there is a substantial difference in base salaries between men and women**. In simpler terms, it suggests that the variations we observe might occur quite frequently due to random factors, making it challenging to attribute them solely to gender-based salary differences. It's important to note, however, that not finding such a difference in this particular study doesn't conclusively prove that there is absolutely no difference in salaries between genders. It simply means that we haven't found sufficient evidence to confidently assert the presence of a notable disparity.