STATE & COUNTY WORKFORCE WEEKLY WAGE EVALUATION

By Gabriel Barela and Jesse St. John

RESEARCH QUESTION:

How does the weekly wage affect the cost of building a single design industrial complex for a budgeted construction contract with a suitable workforce located in the United States?

Real world applications:

Infrastructure Construction

Budget Modeling

WHY IS THIS IMPORTANT?

Main Areas of Interest

- Look at the data at state and county levels for the following:
  - Average weekly wage
  - Employee level (how many workers)

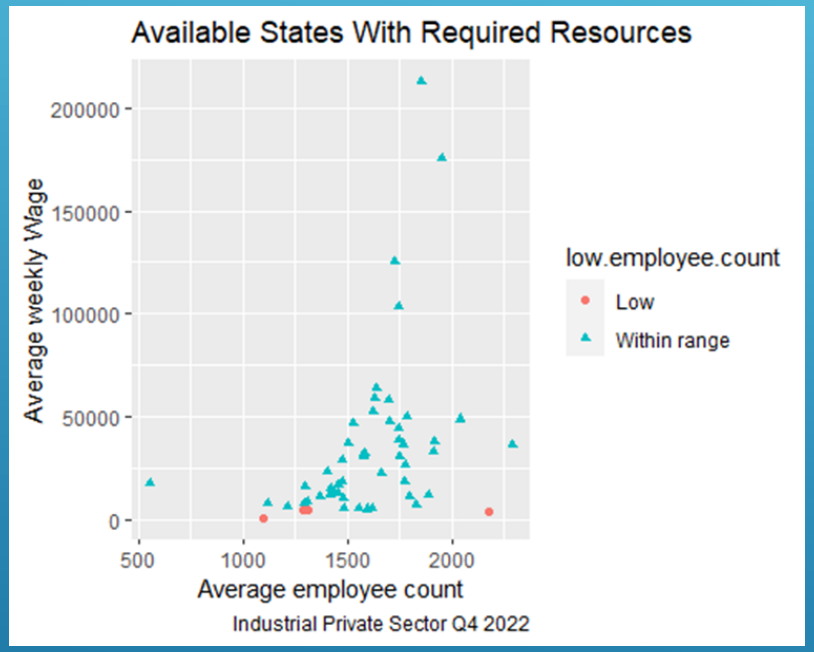
DATA SOURCE:

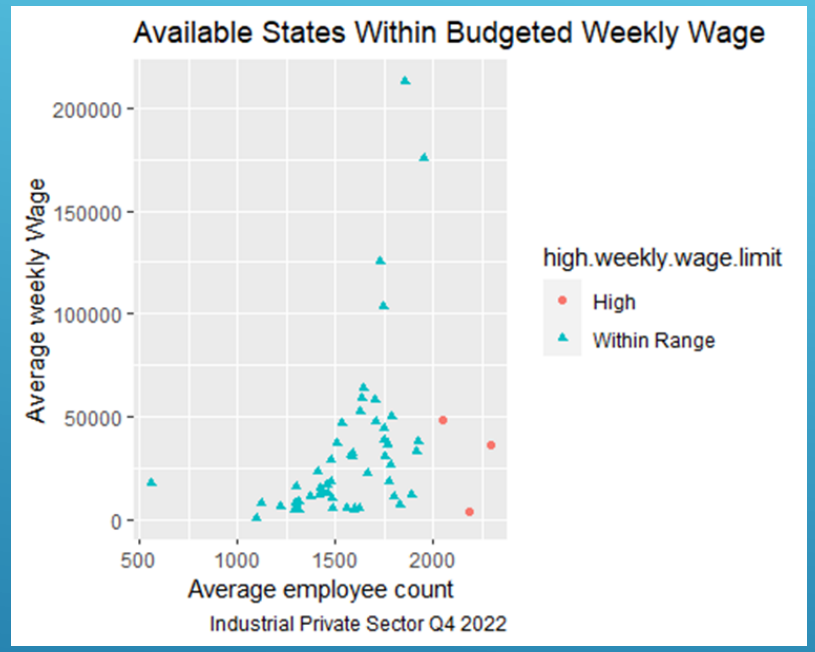
Description: Private, NAICS 236 Construction of buildings, All Counties 2022 Fourth Quarter, All establishment sizes

Source: Quarterly Census of Employment and Wages – Bureau of Labor Statistics

Link: http://www.bls.gov/cew/data/api/2022/4/industry/236.csv

FIRST LOOK AT STATE DATA





Identify locations that do not fit within our specified budget for wage analysis and required workforce resources.

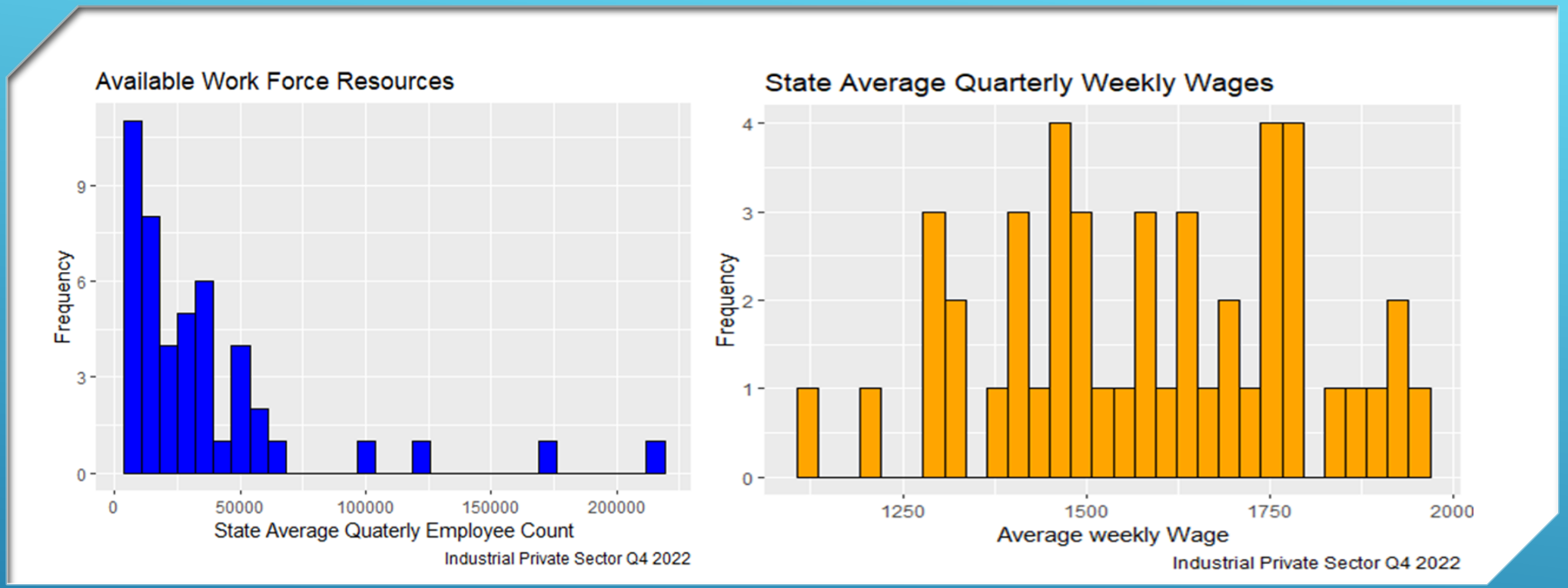
DATA PREPARATION

Pre-R Clean Up

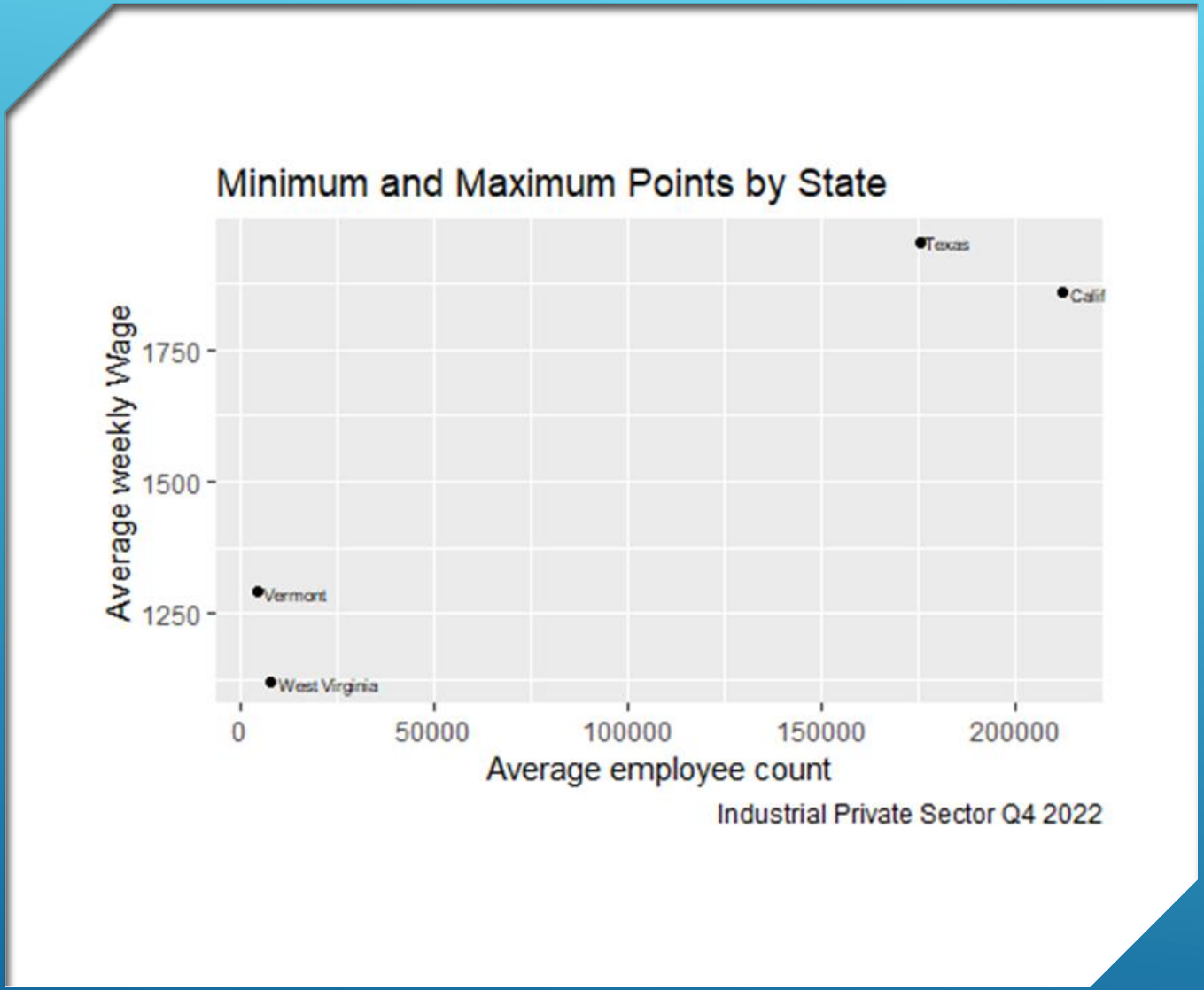
- Reformat of Excel data
  - Convert codes to strings
- Decide which information to focus on
  - Ignore data set internal calculated values such as Location quotient
  - Analyze actual numbers by location

R Clean Up

- Get rid of data from non-private companies
- Ignore sources with an "N" disclosure code
- Calculate average quarterly employee level
- Only include sources with an average quarterly employee level > 4000 at the state level and > 1000 at the county level
- We want an average weekly wage < 2000
- Remove locations outside of the continental United States.

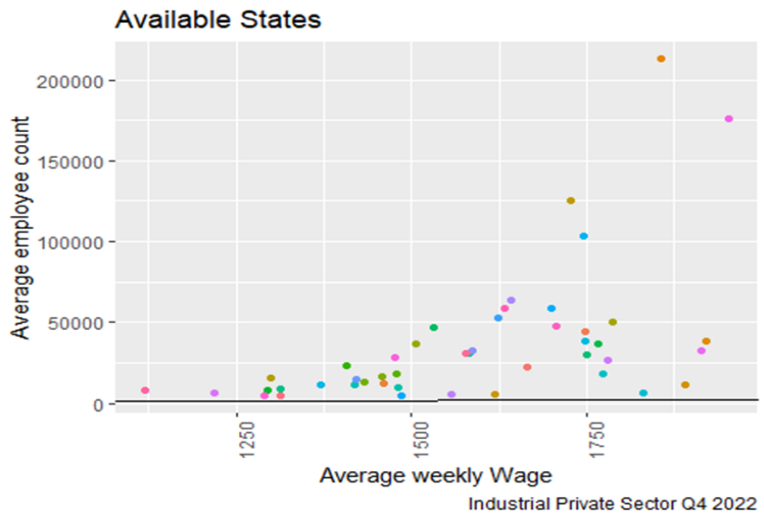


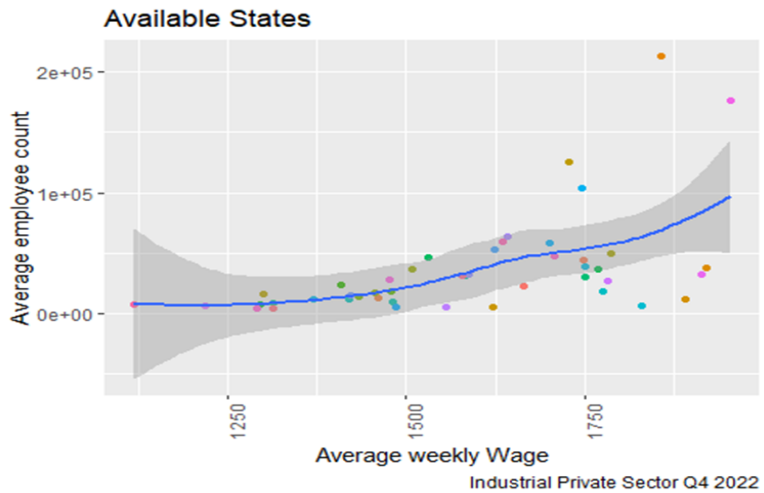
STATE FREQUENCIES AFTER CLEAN UP



IDENTIFY MIN AND MAX

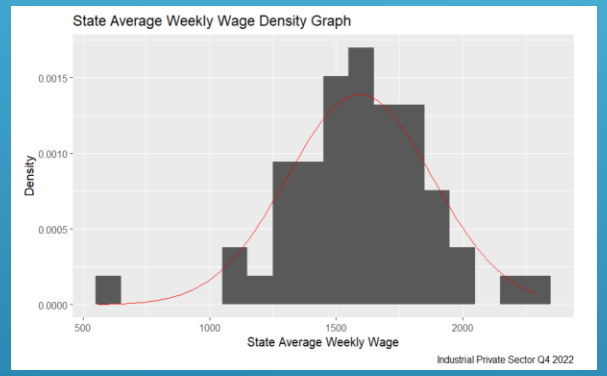
- Examine MIN and MAX for both variables
  - Drive our model from Normal distribution
- Weekly Wage more Normally distributed
- Average Employee Level is not Normally distributed

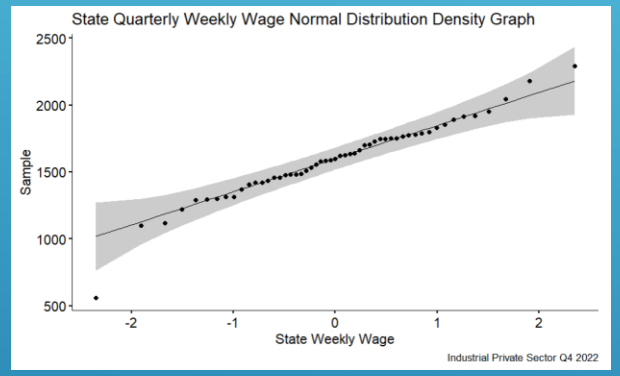


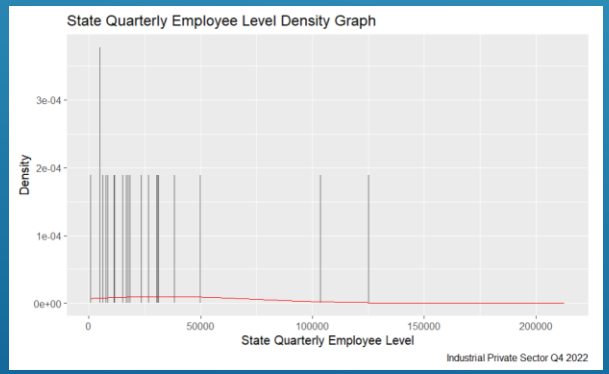


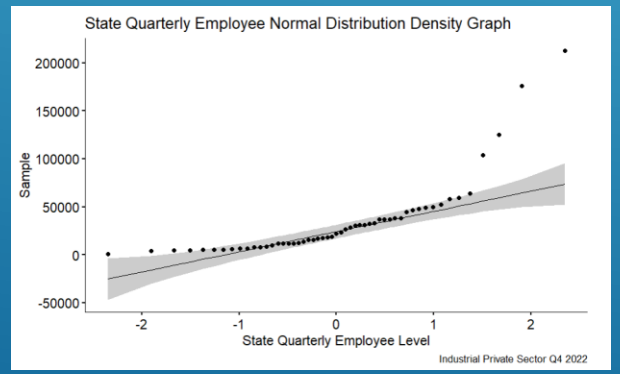
STATE LINEARITY ASSESSMENT

STATE NORMAL DISTRIBUTION COMPARISON









- We still have some states that are within our model parameters, but outside of the normal distribution span
- Our model still has a large range after location reduction
- With the visual comparison no assumptions can be made about the correlation of the two variables.
- Further evaluation is required to assess variance by their means

MODEL HIGHLIGHTS

Count Data

- Whole numbers
- How often something does *not* happen is often not known
- Examine frequencies

Why Use a Poisson Errors Model?

Conventional linear regression methods are not appropriate for use with Count Data:

- Linear models may lead to negative prediction counts
- Errors are not normally distributed
- Variance of the response value is likely to increase with the mean
- 0's can be a problem for data transformations

Null Hypothesis

The average quarterly weekly wage is not affected by the size of the employee level availability by location.

Alternative Hypothesis

The average quarterly weekly wage is affected by the size of employee level availability by location.

METHODS DESCRIPTION

Poisson Errors Model

- Uses the log link, ensures all fitted values are positive
- Poisson Errors account for the integer data with variances that are equal to their means
- Deviance calculation:

Model	Deviance	Error	Link
linear	$\sum (y - \hat{y})^2$	Gaussian	identity
log linear	$2\sum y \log\left(\frac{y}{\hat{y}}\right)$	Poisson	log

ANOVA

- Used when the explanatory variables are categorical
- One-way ANOVA examines if there are statistical differences between the means of three or more independent groups
- For our study:
  - Assume average quarterly employee level per state/county is independent
  - Is there a difference or is it due to chance?

Poisson Error

```
## Call:
## glm(formula = dat.private.state$avg_wkly_wage ~ dat.private.state$avg_qtrly_emplv,
##     family = poisson)
##
## Coefficients:
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept)                   7.312e+00  4.884e-03 1497.17   <2e-16 ***
## dat.private.state$avg_qtrly_emplv 1.492e-06  8.143e-08   18.33   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 1208.53  on 45  degrees of freedom
## Residual deviance:  890.27  on 44  degrees of freedom
## AIC: 1317.3
##
## Number of Fisher Scoring iterations: 4
```

- The residual deviance (890.27) is much larger than the residual degrees of freedom (44).
- This indicates we have overdispersion in this model. This is confirmed by a p-value < .05.
- We can compensate for the overdispersion in our data by attempting to refit the model using quasipoisson errors.

STATE LEVEL REGRESSION MODEL

Quasipoisson Error

```
## Call:
## glm(formula = dat.private.state$avg_wkly_wage ~
dat.private.state$avg_qtrly_emplv,
##     family = quasipoisson)
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     7.312e+00  2.199e-02  332.48  < 2e-16
***
## dat.private.state$avg_qtrly_emplv 1.492e-06  3.667e-
07    4.07 0.000192 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be
20.27765)
##
##     Null deviance: 1208.53  on 45  degrees of freedom
## Residual deviance:  890.27  on 44  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 4
model2$coefficients
##                   (Intercept) dat.private.state$avg_qtrly_emplv
##                  7.311596e+00                      1.492468e-06
```

- The residual deviance (890.27) and the residual degrees of freedom (44) did not change with the new model.
- This indicates we still have over dispersion in our data.
- The p value increased, but it is still less than the desired value of .05. Thus our model is still not a good fit for our data.

STATE LEVEL REGRESSION MODEL

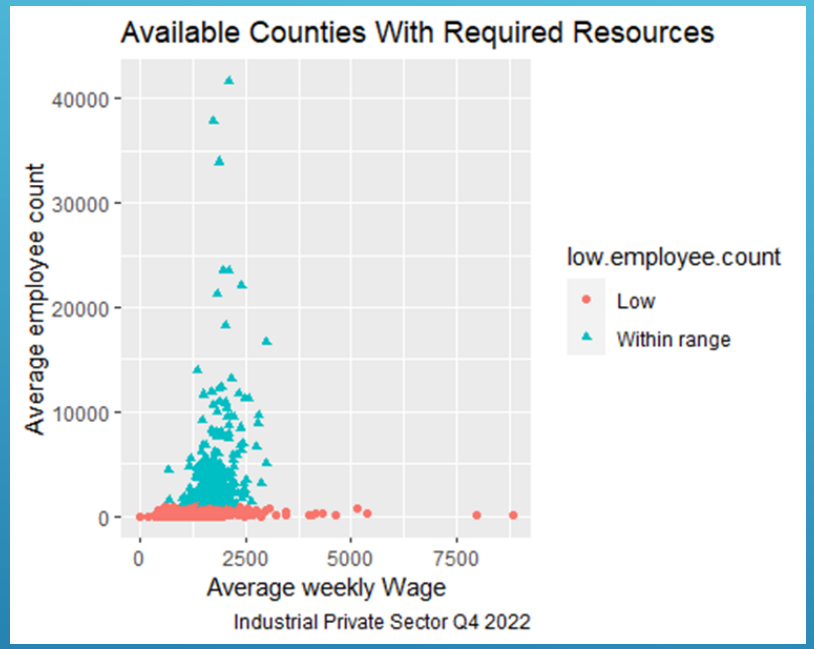
STATE LEVEL ANOVA

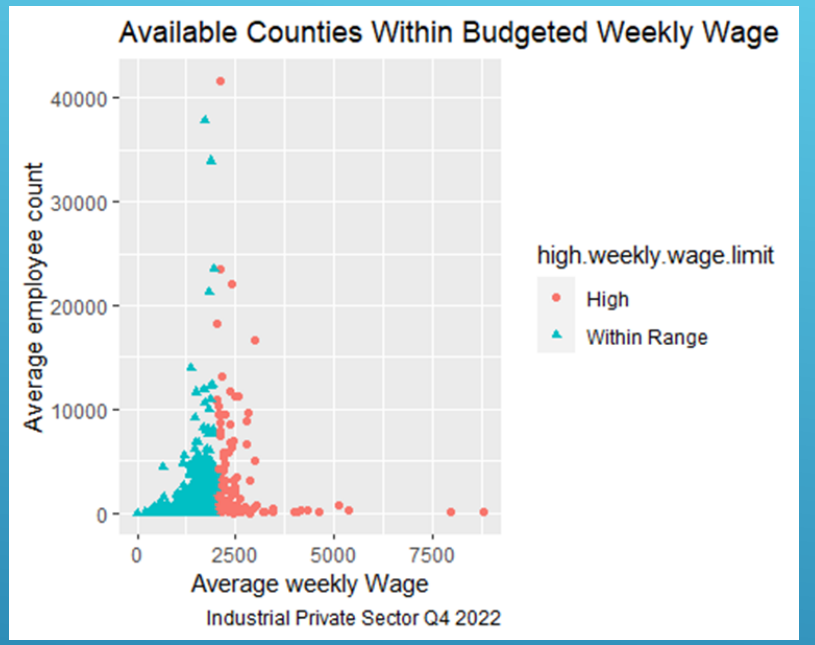
- F test statistic of 17.29 is greater than the critical value of F = 4.04.
- Thus we can reject our null hypothesis and conclude that wage is affected by employee level.
- The Shapiro-Wilk normality test was done to verify whether or not our residuals come from a normal distribution.
- Since our p-value for this test is greater than .05, we do not reject the null hypothesis, and thus the residual data do follow a normal distribution.

```
## Call:
## aov(formula = avg_wkly_wage ~ avg_qtrly_emplvl, data =
dat.private.state)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -390.85  -98.94  -12.41  116.23  369.25
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.491e+03  3.423e+01  43.556  < 2e-16 ***
## avg_qtrly_emplvl 2.570e-03  6.179e-04   4.158 0.000146 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 175.6 on 44 degrees of freedom
## Multiple R-squared:  0.2821,    Adjusted R-squared:  0.2658
## F-statistic: 17.29 on 1 and 44 DF,  p-value: 0.0001459

##qf(0.95, 1, 48)
## [1] 4.042652
##shapiro.test(one.way$residuals) #will probably not need
##
##  Shapiro-Wilk normality test
 ##
## data:  one.way$residuals
 ## W = 0.97829, p-value = 0.5377
```

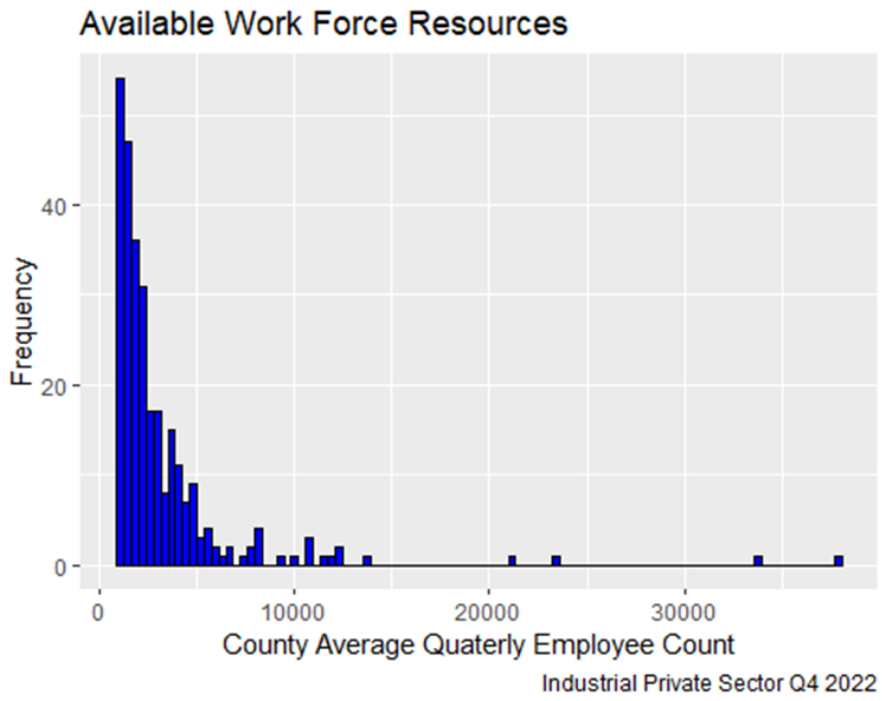
FIRST LOOK AT COUNTY DATA

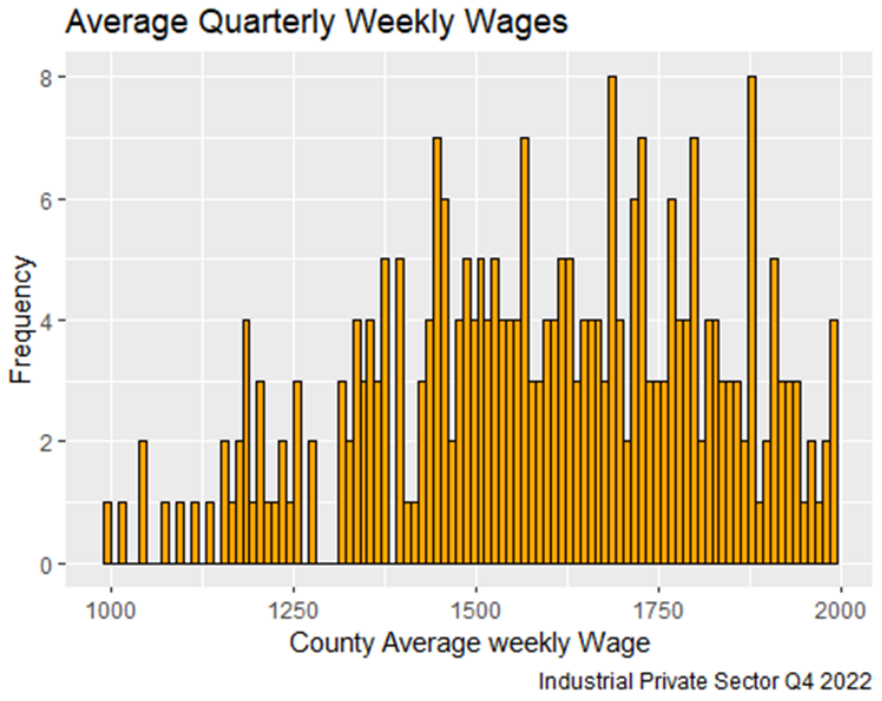




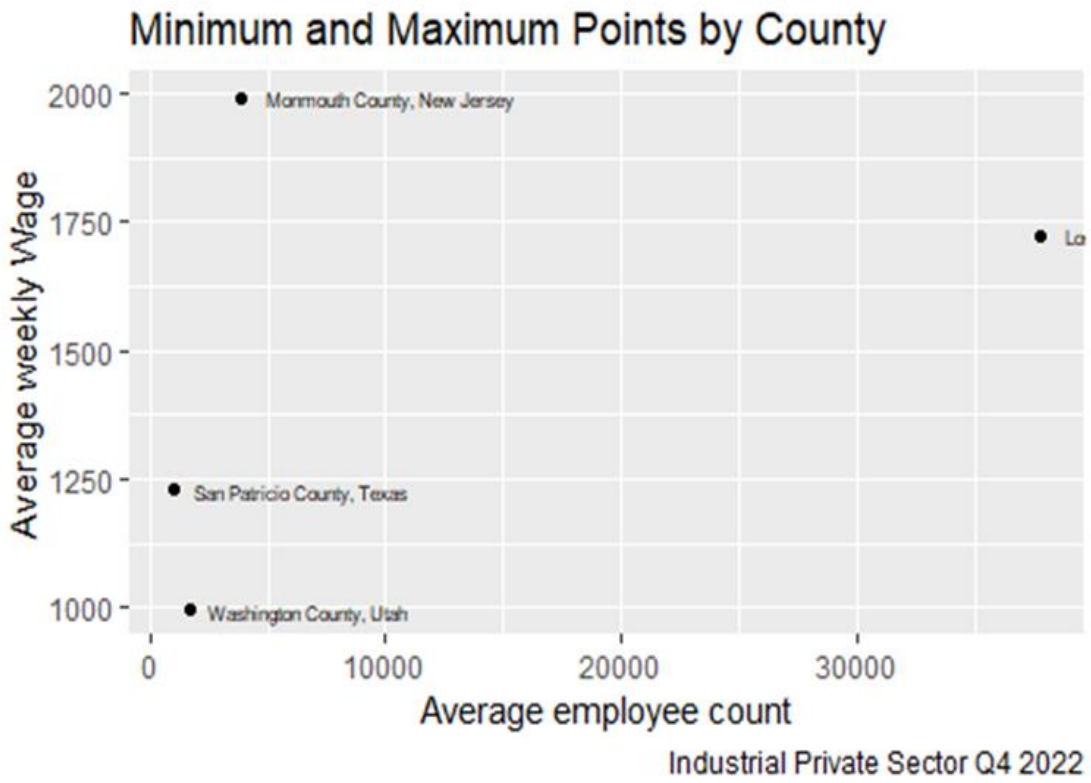
Identify locations that do not fit within our specified budget for wage analysis and required workforce resources.

CLEAN COUNTY DATA UP IN THE SAME WAY AS THE STATE LEVEL



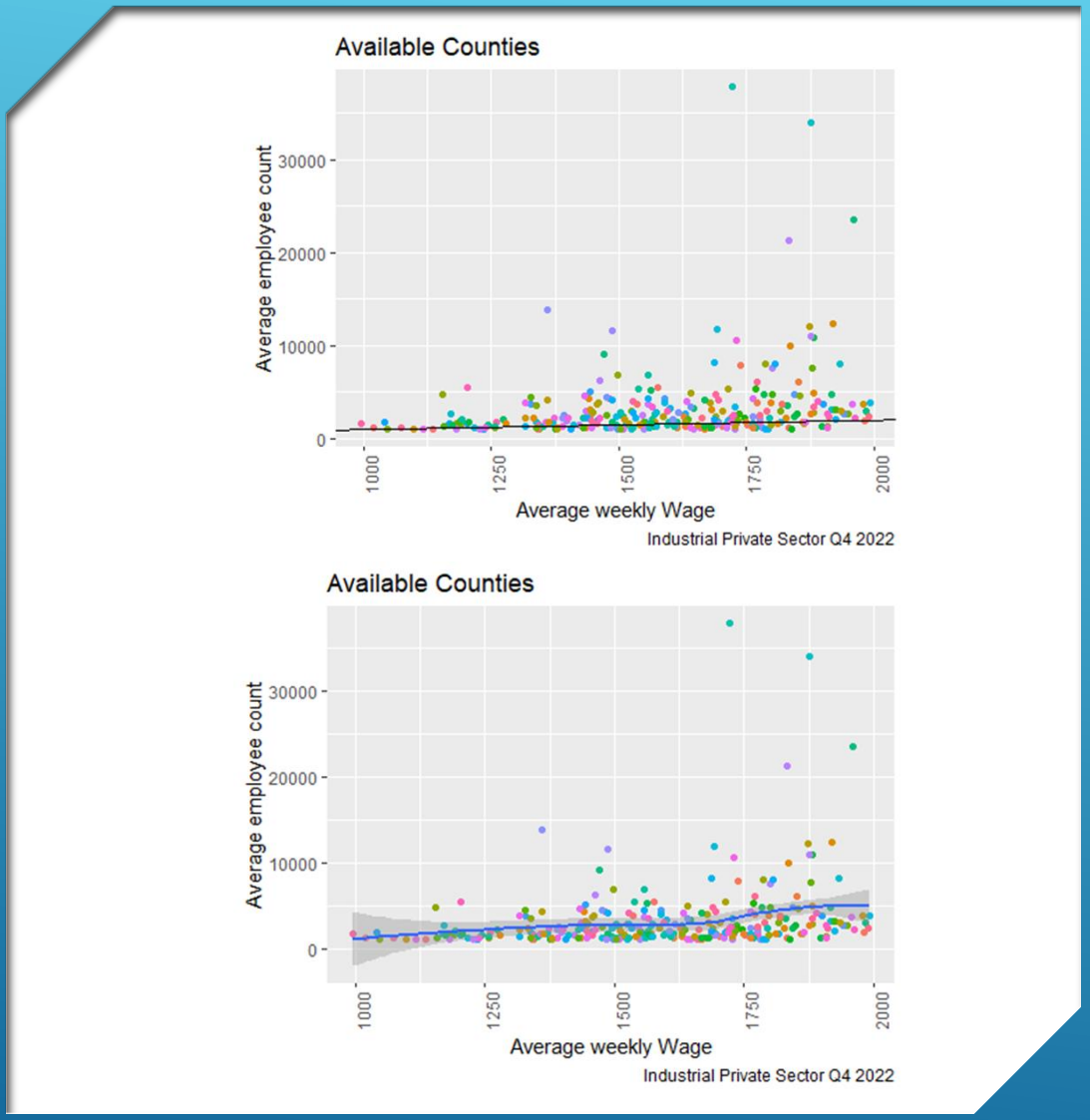


COUNTY FREQUENCIES AFTER CLEAN UP



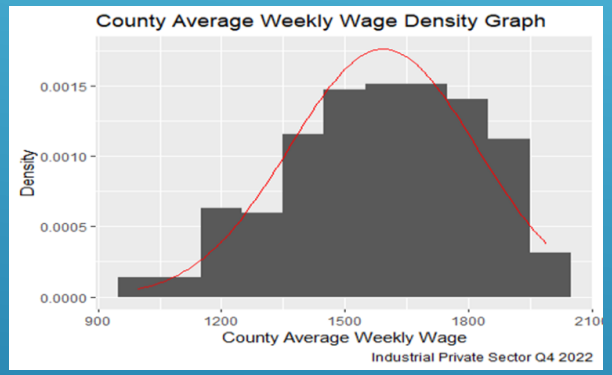
IDENTIFY MIN AND MAX

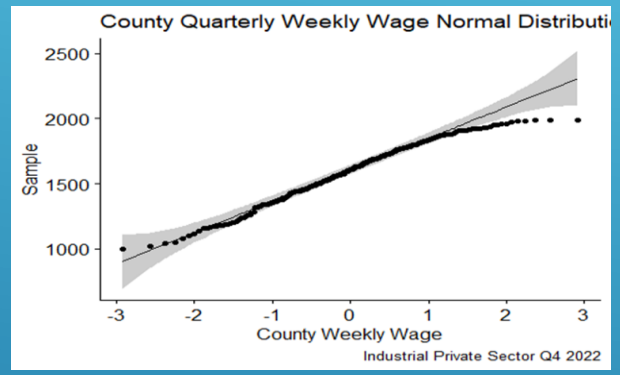
- Examine MIN and MAX for both variables
  - Drive our model from Normal distribution
- Weekly Wage more Normally distributed
- Average Employee Level is not Normally distributed

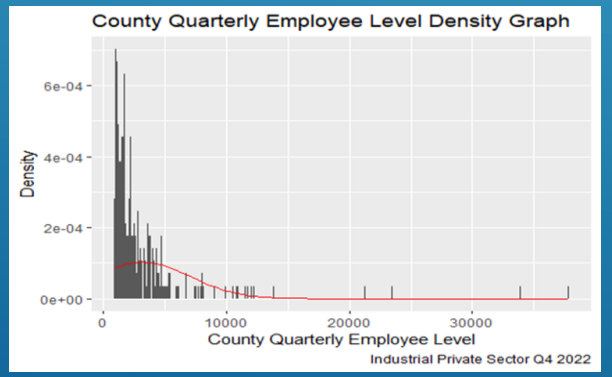


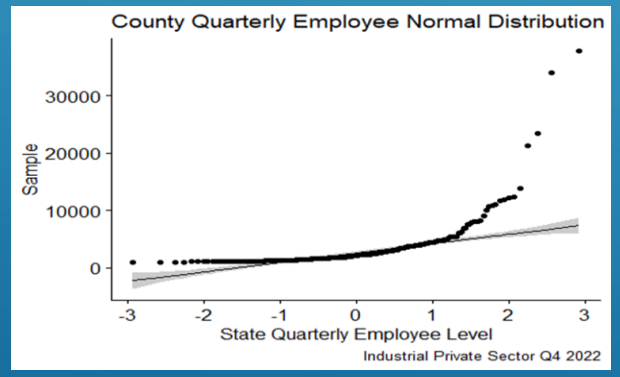
COUNTY LINEARITY ASSESSMENT

COUNTY NORMAL DISTRIBUTION COMPARISON









- We still have some states that are within our model parameters, but outside of the normal distribution span
- Our model still has a large range after location reduction
- With the visual comparison no assumptions can be made about the correlation of the two variables.
- Further evaluation is required to assess variance by their means

Poisson Error

```
## Call:
##  glm(formula = dat.private.county$avg_wkly_wage ~
dat.private.county$avg_qtrly_emplv,
##      family = poisson)
##
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)                     7.347e+00  1.893e-03 3881.18   <2e-16 ***
## dat.private.county$avg_qtrly_emplv 8.033e-06  3.500e-07   22.95   <2e-16
***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 9405.0  on 284  degrees of freedom

## Residual deviance: 8907.6  on 283  degrees of freedom
##  AIC: 11534
##
## Number of Fisher Scoring iterations: 4

model3$coefficients

##                  (Intercept) dat.private.county$avg_qtrly_emplv
##                 7.346851e+00                      8.033384e-06
```

- The residual deviance (8907.6) is much larger than the residual degrees of freedom (283).
- This indicates we have overdispersion in this model as well. Once again confirmed by a p-value < .05.
- We can compensate for the overdispersion in our data by attempting to refit the model using quasipoisson errors.

COUNTY LEVEL REGRESSION MODEL

Quasipoisson Error

```
## Call:
## glm(formula = dat.private.county$avg_wkly_wage ~
dat.private.county$avg_qtrly_emplv,
##     family = quasipoisson)
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     7.347e+00  1.050e-02 699.866  < 2e-16 ***
## dat.private.county$avg_qtrly_emplv 8.033e-06  1.941e-06   4.139
4.61e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 30.75371)
##
##     Null deviance: 9405.0  on 284  degrees of freedom
## Residual deviance: 8907.6  on 283  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 4
```

```
model4$coefficients
```

```
##                    (Intercept) dat.private.county$avg_qtrly_emplv
##                   7.346851e+00                     8.033384e-06
```

- The residual deviance (8907.6) and the residual degrees of freedom (351) did not change with the new model.
- This indicates we still have overdispersion in our data.
- While the p-value did increase, it is still much smaller than .05. Therefore, our model at the county level is also probably not a good fit for our data.

COUNTY LEVEL REGRESSION MODEL

COUNTY LEVEL ANOVA

- F test statistic of 17.27 is greater than the critical value of F = 4.04.
- Thus we can reject our null hypothesis and conclude that the average weekly wage is affected by employee level.

```
## Call:
##  aov(formula = avg_wkly_wage ~ avg_qtrly_emplvl, data =
dat.private.county)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -574.30 -148.04   24.36  167.30  407.40
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
##
## (Intercept)     1.548e+03  1.697e+01  91.228  < 2e-16 ***
## avg_qtrly_emplvl 1.395e-02  3.358e-03   4.156  4.3e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 220.5 on 283 degrees of freedom
## Multiple R-squared:  0.05751,    Adjusted R-squared:  0.05418
## F-statistic: 17.27 on 1 and 283 DF,  p-value: 4.302e-05

qf(0.95, 1, 48)

## [1] 4.042652
```

MODEL EVALUATION

Methods Used

- Poisson and quasipossion regression
- One-way ANOVA

Data Analysis

- Reject null hypothesis at both levels

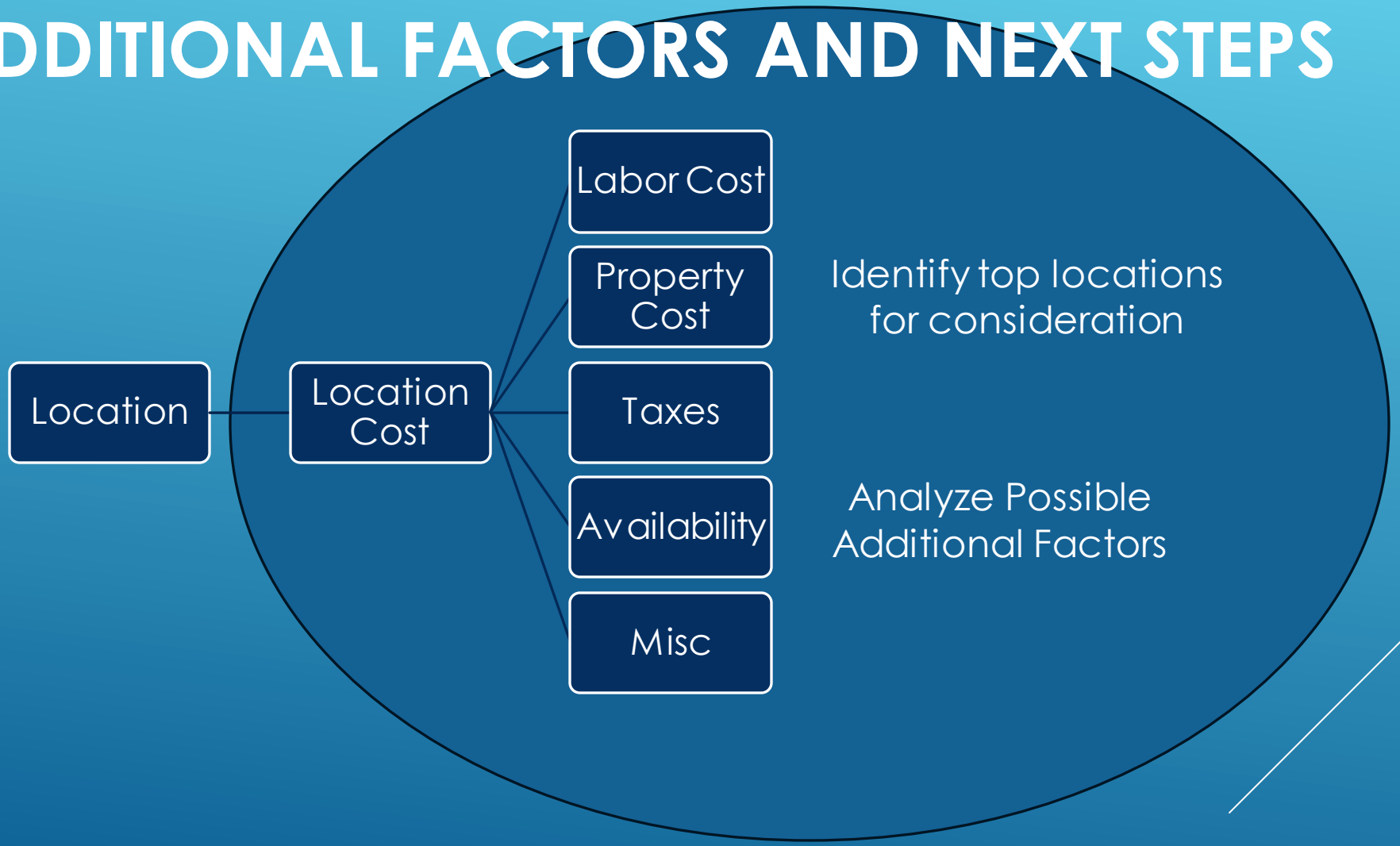
Cautionary

- Over dispersion
  - Feasible that employee level availability affects weekly wage
  - Other factors should be evaluated

Conclusion

- Identifies Suitable locations for deeper analysis
- Allows for evaluation additional factors

ADDITIONAL FACTORS AND NEXT STEPS



REFERENCES

Crawley, M. J. Statistics: An introduction using R (2nd ed.), ISBN: 978-1-118-94109-6.