36402 Data Exam 2

Joong Ho Choi

TOTAL POINTS

94 / 100

QUESTION 1

1 Introduction and Format 17 / 20

- 0 pts All items are excellent (Contents, Figures & Tables, Questions & Objectives, Final Findings).

Overall Requirements: Contents

✓ **- 3 pts** **R codes/outputs are included**

- 5 pts Not a report. Simply a numbered or bulleted list of responses to the questions.

- 1 pts Includes mathematical derivations

- 1 pts Very divergent format (different line spacing, margins, and/or font size).

Overall Requirements: Format

- 1 pts Some unlabeled/mistakenly labeled figures/tables.

- 2 pts Figures/tables unclear.

- 2 pts Figures/tables without captions.

Introduction: Questions & Objectives

- 1 pts All goals mentioned but are vague.

- 2 pts Maybe you can add more background introductions.

- 2 pts Not all objectives from the instructions are explicitly mentioned upfront

- 4 pts Questions and objectives not addressed

Introduction: Final Findings

- 1 pts Final findings unclear/not sufficient.

- 3 pts Final findings not reported

QUESTION 2

2 Exploratory Data Analysis 15 / 15

✓ **- 0 pts** **Correct**

Data Description

- 0 pts Correct

- 1 pts Missing meaning of relevant variables given with units

Response Variable

- 0 pts Correct

- 3 pts Missing

- 2 pts Missing distribution of the response variable

Univariate EDA

- 0 pts Correct

- 3 pts Missing

- 1.5 pts Missing interpretation

- 2 pts Missing graphs

- 2 pts Some results (e.g. average, sd, shape) are not connected to the modeling decisions or results

Multivariate EDA

- 0 pts Correct

- 3 pts Missing

- 1 pts It's better to say something here, such as what transformations they might need or what variables appear to be the strongest predictors

- 2 pts No graphs about the predictors and the response variable.

- 0.5 pts Scatterplots are a little unclear

- 1 pts Scatterplots may provide more information than correlation plots since we can observe nonlinear associations from scatterplots, which probably would help us build our models.

- 1.5 pts No graphs about the predictors and the response variable other than LoanAmount

- 1.5 pts Missing interpretation

Trends and Interesting Features

- 0 pts Correct

- 1.5 pts Missing

- 1 pts Wrong analysis on the trend

Relevant Plots

- **0.5 pts** Some irrelevant plots
- **1 pts** More data visualization needed
- **3 pts** No Plots

QUESTION 3

3 Modeling & Diagnostics 27 / 30

Model 1: Poisson regression on `CreateJob`

- **0 pts** Correct
- **2 pts** Misspecify the distribution of the response variable.

✓ - **1 pts** **Not enough explanation for the response distribution.**

- **2 pts** Missing explanation for substantial covariates
- **2 pts** Inappropriate/Missing model description
- **1 pts** Minor errors in model description

Model 2: Poisson Additive Model

✓ - **0 pts** **Correct**

- **2 pts** Misspecify the distribution of the response variable.
- **2 pts** Missing explanation for substantial covariates
- **2 pts** Inappropriate/Missing model description
- **1 pts** Minor errors in model description

Model Diagnostics

- **0 pts** Good enough
- **2 pts** **Model 1:** Missing diagnostics for the response distribution (Poisson). At least a discussion on residuals/responses of fit is required.
- **1 pts** $$ \llcorner $$ **Model 1:** Appropriate plots but with incorrect diagnostics; Poisson regression does not impose the normality/constant variance assumption on errors.
- **2 pts** **Model 2:** Missing diagnostics for the response distribution (Poisson). At least a discussion on residuals/responses is required.
- **1 pts** $$ \llcorner $$ **Model 2:** Appropriate plots but with incorrect diagnostics; Poisson additive model does not impose the normality/constant variance assumption on errors.

✓ - **2 pts** **Model 1&2:** **Missing diagnostics for the model bias: whether a deterministic formula for $$g(\mathbb{E}(Y|X))$$ as a function of the covariates is reasonable. At least a discussion on either of the followings is required: partial residuals/response plot, fitted values vs. true responses, good-ness of fit**

- **1 pts** Other minor errors

Hypothesis Testing via bootstrap

✓ - **0 pts** **Correct**

- **3 pts** Missing/Incorrect choice of Bootstrap method. Semi-/Non-parametric bootstrap is not appropriate here.
- **1 pts** Missing/Inappropriate justification for the bootstrap choice; Parametric bootstrap is justifiable since we are interested in the distribution of our statistics under the null hypothesis.
- **1 pts** **Parametric bootstrap**: Missing comment on what model to use for parameter estimation.
- **1 pts** Inappropriate test statistic
- **1 pts** Incorrect/Missing formulation of the p-value

Confidence Intervals for coefficients via bootstrap

✓ - **0 pts** **Correct**

- **3 pts** Missing justification for the bootstrap choice. Here, semi-/non-parametric bootstrap is good to use.
- **1 pts** Not enough explanation on CI construction
- **2 pts** Missing explanation on CI construction
- **1 pts** Other minor errors
- **6 pts** Missing

QUESTION 4

4 Results 20 / 20

Results for test for nonlinearity

✓ - **0 pts** **Correct**

- **3 pts** The $$p$$-value is not close to zero (as it should be).
- **3 pts** No plot demonstrates the shape of the relationship between loan amount and job creation.
- **10 pts** Missing

Confidence intervals

✓ **- 0 pts Correct**

- 5 pts No table presents CIs for association between industry and jobs created.

- 3 pts Insufficient commentary on the CIs (e.g. no statement on what the CIs mean and no mention of which industries have the highest and lowest rates).

- 5 pts No comments on the CIs.

- 10 pts Missing

QUESTION 5

5 Conclusions 15 / 15

✓ **- 0 pts Excellent! (Style, Findings and Discussion)**

Overall Requirements: Writing & Style

- 1 pts Too much stats jargon.

- 3 pts Is hard to understand for non-statisticians

Conclusions: Main Findings

- 0 pts Perfect.

- 2 pts No discussion on different industries.

- 2 pts No discussion on (non-)linearity

- 3 pts Partially missing

- 5 pts Missing

Conclusions: Discussion & Limitations

- 0 pts Excelent.

- 2 pts You should mention that this was not a causal analysis.

- 2 pts not enough discussion on limitations

- 5 pts Missing

- 1 pts Do not write out a list!

- 1 pts No labels for answers

- 15 pts Missing

36-402 DA Exam 2

Joong Ho Choi (joonghoc)

April 29, 2022

Introduction

The United States Small Business Administration (SBA) is an agency of the US federal government intended to help small businesses, and hence create jobs and boost the economy. One way it does this is through the SBA 7(a) Loan Program. Banks are sometimes reluctant to give loans to small businesses—which might use the loans to buy equipment or hire staff—because small businesses are risky and may not pay back the loans. The loan program lets the federal government take some of the risk: Businesses apply to the SBA for a loan guarantee and to a bank for a loan; if the applications are approved, the SBA will guarantee the loan, meaning that if the business fails to pay it back, the SBA will pay the bank. Considering the risks that come with loans, it is essential that they understand the relationship between jobs created and dollars loanded, and that they provide loans to the types of businesses most likely to create jobs. Thus, we will investigate whether whether the relationship between jobs created and dollars loaned is linear **1** and what kinds of businesses create the most jobs per dollar loaned **2**.

Our final findings were that the relationship between jobs created and dollars loaned is non-linear. A loan that's twice as big will likely support fewer than twice as many jobs. **1** Also, businesses of public admin industry create the most jobs per dollar loaned, whereas businesses of construction industry create the least. **2**

Exploratory Data Analysis

Our dataset contains data on 16,245 SBA loans granted to businesses in Pennsylvania between 1995 and 2014. Relevant variables for our investigation are the followings:

1 Introduction and Format 17 / 20

- **0 pts** All items are excellent (Contents, Figures & Tables, Questions & Objectives, Final Findings).

Overall Requirements: Contents

✓ - **3 pts** **R codes/outputs are included**

- **5 pts** Not a report. Simply a numbered or bulleted list of responses to the questions.
- **1 pts** Includes mathematical derivations
- **1 pts** Very divergent format (different line spacing, margins, and/or font size).

Overall Requirements: Format

- **1 pts** Some unlabeled/mistakenly labeled figures/tables.
- **2 pts** Figures/tables unclear.
- **2 pts** Figures/tables without captions.

Introduction: Questions & Objectives

- **1 pts** All goals mentioned but are vague.
- **2 pts** Maybe you can add more background introductions.
- **2 pts** Not all objectives from the instructions are explicitly mentioned upfront
- **4 pts** Questions and objectives not addressed

Introduction: Final Findings

- **1 pts** Final findings unclear/not sufficient.
- **3 pts** Final findings not reported

36-402 DA Exam 2

Joong Ho Choi (joonghoc)

April 29, 2022

Introduction

The United States Small Business Administration (SBA) is an agency of the US federal government intended to help small businesses, and hence create jobs and boost the economy. One way it does this is through the SBA 7(a) Loan Program. Banks are sometimes reluctant to give loans to small businesses—which might use the loans to buy equipment or hire staff—because small businesses are risky and may not pay back the loans. The loan program lets the federal government take some of the risk: Businesses apply to the SBA for a loan guarantee and to a bank for a loan; if the applications are approved, the SBA will guarantee the loan, meaning that if the business fails to pay it back, the SBA will pay the bank. Considering the risks that come with loans, it is essential that they understand the relationship between jobs created and dollars loanded, and that they provide loans to the types of businesses most likely to create jobs. Thus, we will investigate whether whether the relationship between jobs created and dollars loaned is linear **1** and what kinds of businesses create the most jobs per dollar loaned **2**.

Our final findings were that the relationship between jobs created and dollars loaned is non-linear. A loan that's twice as big will likely support fewer than twice as many jobs. **1** Also, businesses of public admin industry create the most jobs per dollar loaned, whereas businesses of construction industry create the least. **2**

Exploratory Data Analysis

Our dataset contains data on 16,245 SBA loans granted to businesses in Pennsylvania between 1995 and 2014. Relevant variables for our investigation are the followings:

-CreateJob: Number of jobs the business expects to create using the loan money

-UrbanRural Whether the business is in an urban or rural area

-NewBusiness: Whether the business is new or already existed

-Franchise: Y if the business is a franchise (part of a chain, like McDonald's) or N if the business is independent

-DisbursementGross: Total amount of money loaned

-shortdesc: Description of the industry category the business is in, such as "Construction" or "Finance and Insurance". description gives the full description, shortdesc is a short version suitable for use as a factor variable in models.

There are other variables in the dataset, but they are not important for our analysis.

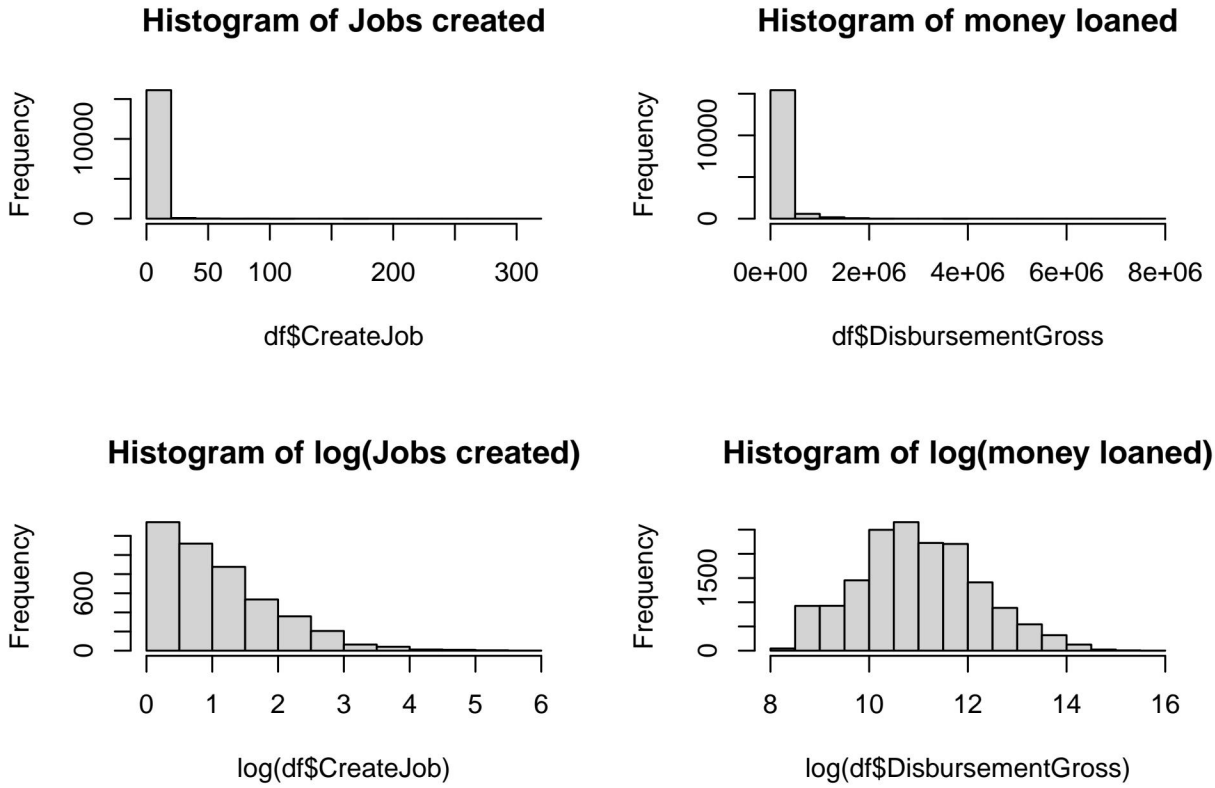


Figure 1: Histograms showing the distributions of variables CreateJob and Disbursement

We began by examining the univariate distributions of the key continuous variables(DisbursementGross as predictor and CreateJob as response).While the number of jobs created ranged from 0 to 320 with mean of 1.387935, the money loaned ranged from 4000 to 7699233 with the mean of 132291.9. Visually, it was clear to see that

the histograms for both variables seem to be heavily skewed to the right. Even after attempting log transformation, the distribution of Jobs created was still heavily skewed to the right,whereas the histogram of log(money loaned) became more symmetrical. Thus, this suggested to me that Jobs created likely follow poisson distribution.

For other key variables that are categorical, we used barplots and table format.

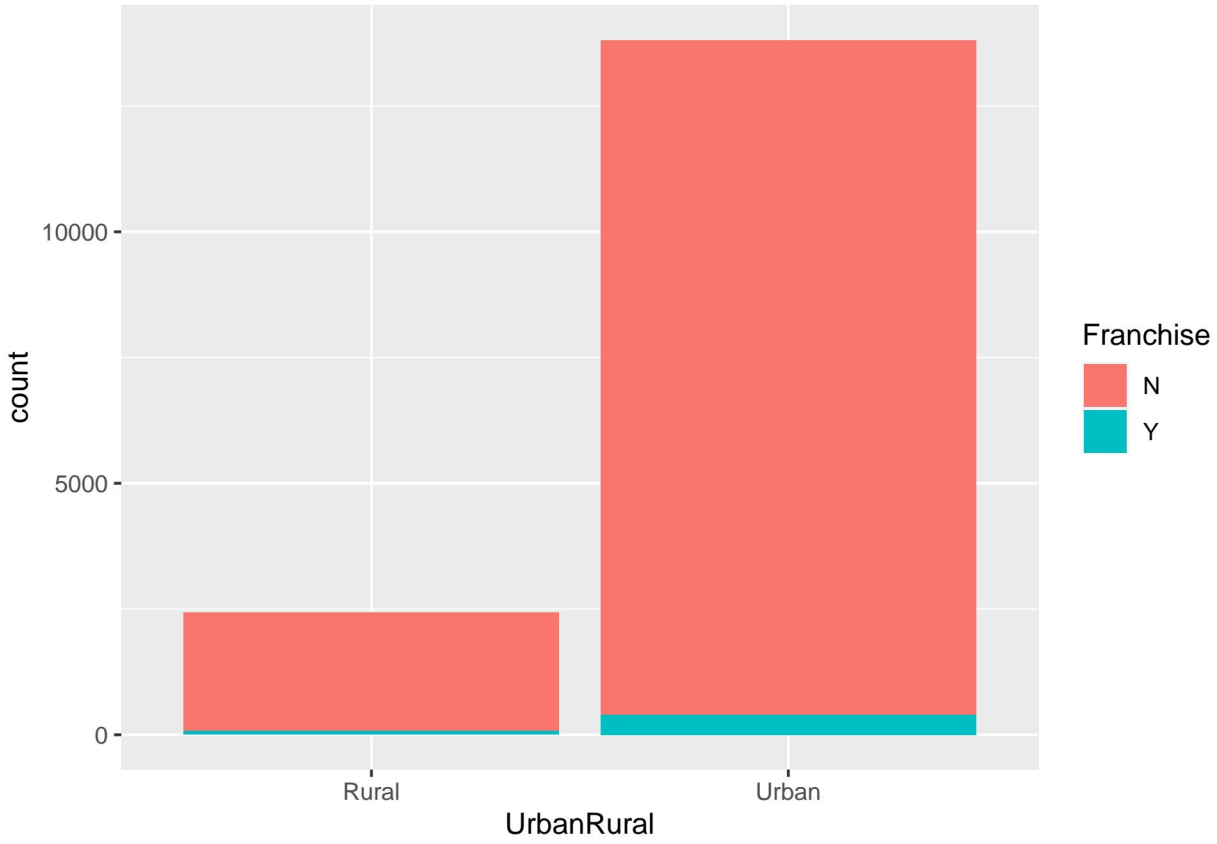


Figure 2: Barplot of Franchise given UrbanRural

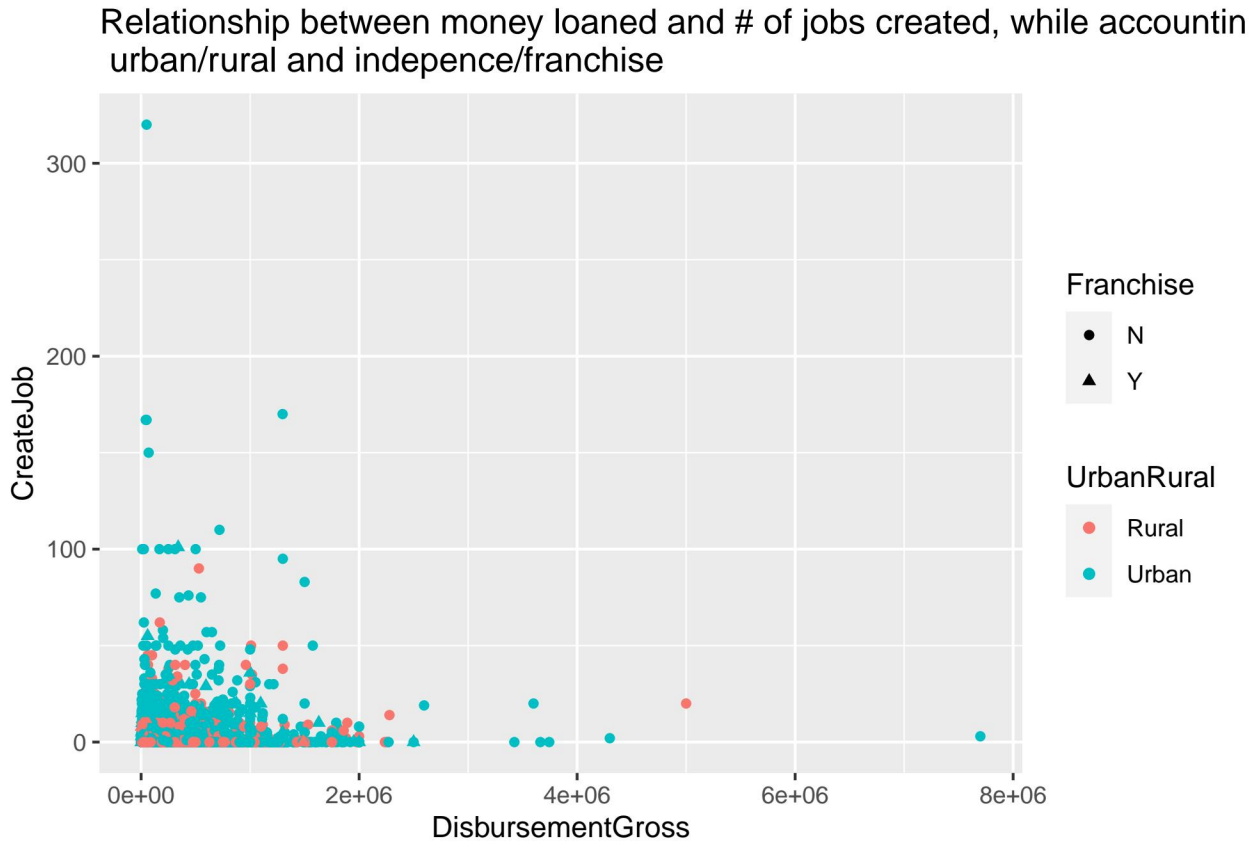
From fig 2 above, we see there are more urban firms than rural firms. Also, there are more independent firms than Franchise firms. In fact, Independent firms are the majority of firms in both rural and urban areas.

```
##
## AccommodationFood      Construction             Other      ProfServices
##              1574              1654              1754              1944
##       RetailTrade
##              2769
```

Table 1: Frequency table of top 5 industries

Also, in terms of count, these are the top 5 industry categories that businesses are in; I chose to explore only top 5 for data visualization purpose, as there are too many industries to fit into one clear visualization for later. From the table, we see retail trade is the top 1, which consists 17% of the data, and transporation warehousing is the top 5th, which consists 9.68% of the data.

Then, for multivariate visualization, we created a scatter plot to visualize the relationships between continuous predictor and response, and boxplots for relationship between categorical predictors and the response variable.



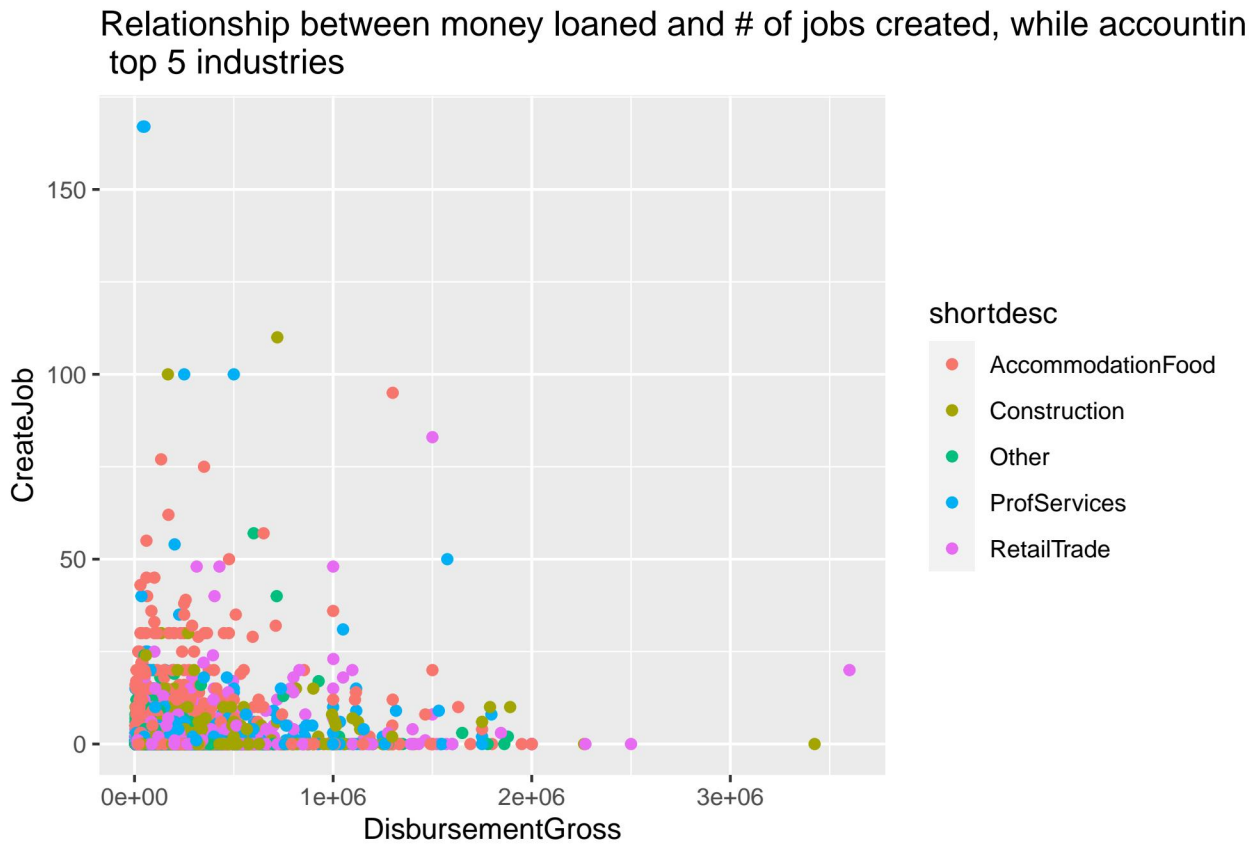


Figure 3: Scatterplots that show relationship between # of jobs created and money loaned, while accounting for other business variables

From both scatterplots, it seems clear that there is no linear relationship between the money loaned and the # of jobs created. It also seems that -# of jobs created is greater for urban businesses than rural businesses -# of jobs created is greater for independent businesses than franchise businesses -among the top 5 industry, # of jobs created is greatest for prof services and least for the other.

Considering how there's no relationship between the rank and the # of jobs created, my models will definitely need to account for the remaining industries as well. In addition, interestingly, heavier points cluster for relatively higher # of jobs created and wider spread for accommodation food and prof service industries suggest that in the model analysis later on, models' coefficients will be higher for Accommodation food and prof services industries than for other 3 industries in top 5.

2 Exploratory Data Analysis 15 / 15

✓ **- 0 pts Correct**

Data Description

- 0 pts Correct

- 1 pts Missing meaning of relevant variables given with units

Response Variable

- 0 pts Correct

- 3 pts Missing

- 2 pts Missing distribution of the response variable

Univariate EDA

- 0 pts Correct

- 3 pts Missing

- 1.5 pts Missing interpretation

- 2 pts Missing graphs

- 2 pts Some results (e.g. average, sd, shape) are not connected to the modeling decisions or results

Multivariate EDA

- 0 pts Correct

- 3 pts Missing

- 1 pts It's better to say something here, such as what transformations they might need or what variables appear to be the strongest predictors

- 2 pts No graphs about the predictors and the response variable.

- 0.5 pts Scatterplots are a little unclear

- 1 pts Scatterplots may provide more information than correlation plots since we can observe nonlinear associations from scatterplots, which probably would help us build our models.

- 1.5 pts No graphs about the predictors and the response variable other than LoanAmount

- 1.5 pts Missing interpretation

Trends and Interesting Features

- 0 pts Correct

- 1.5 pts Missing

- 1 pts Wrong analysis on the trend

Relevant Plots

- 0.5 pts Some irrelevant plots

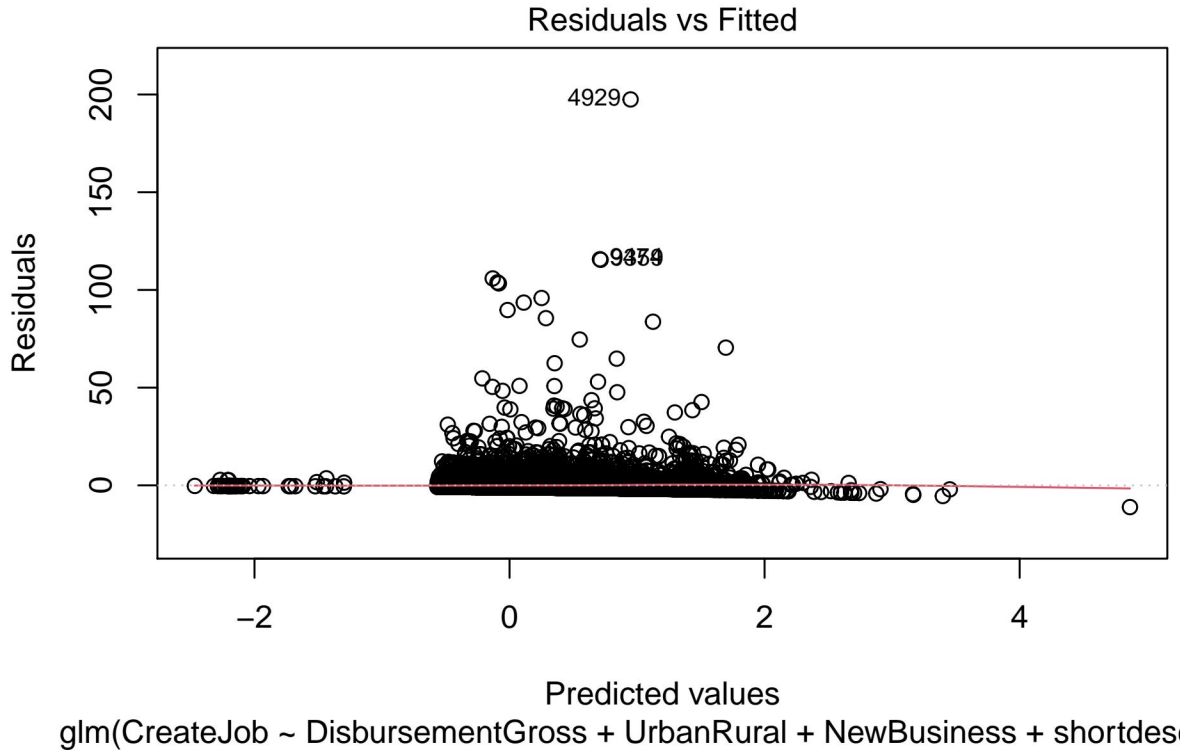
- 1 pts More data visualization needed

- 3 pts No Plots

Modeling & Diagnostics

Four model 1, we fitted a GLM with poisson response distribution to study the relationship between business type and jobs per dollar. To be specfic, this model included in relevant business covariates, such as UrbanRural, NewBusiness, shortdesc, and Franchise. It also included in loan amount(DisbursementGross) for predicting the number of jobs created. However, it's unclear whether the relationship between jobs and loan amount is linear. Thus, to check this, I fitted a model 2(GAM), allowing mgcv to choose the smoothing parameter automatically, up to a maximum edf of 5, for DisbursementGross.

After fitting the two models, I performed residual diagnostics on them.



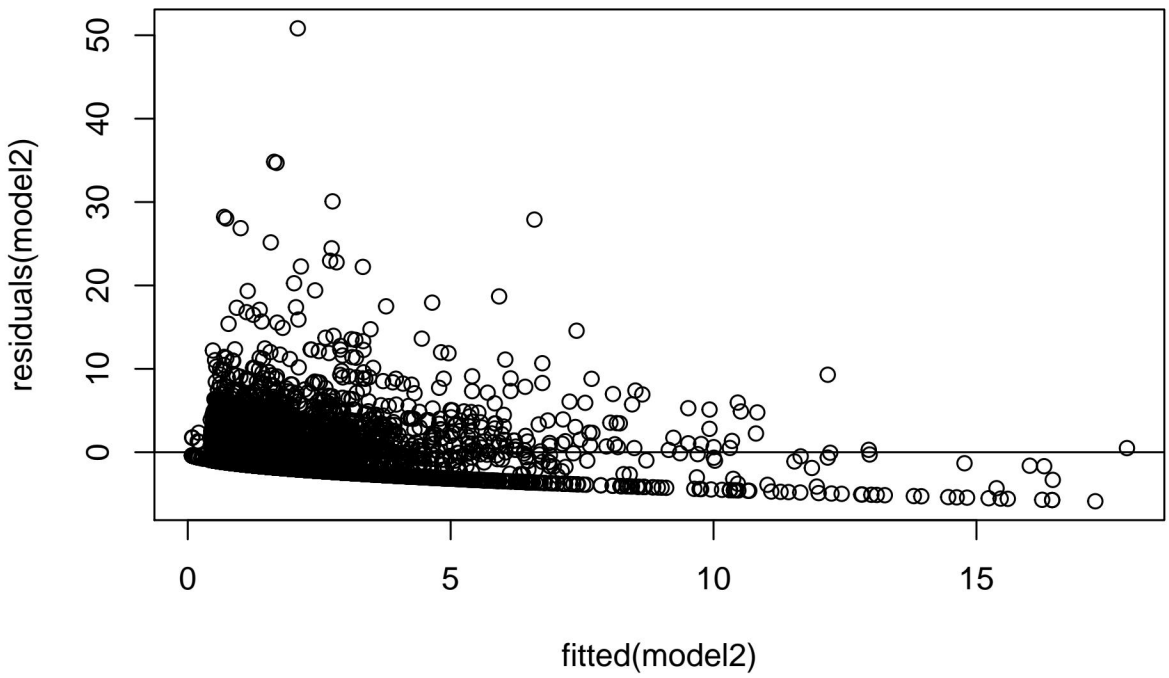


Figure 4: Diagnostic plots analyzing fitted values versus respective residuals for models 1 and 2

From the residual plots for both models, it seems homoscedasticity is greatly violated for both. Is is also worth noting there are many outliers. Heteroscedasticity is a problem because ordinary least squares (OLS) regression assumes that all residuals are drawn from a population that has a constant variance. This might affect the confidence intervals of the two models' predictions to be less reliable, and slightly undermine the validity of each model's ability to perform inference.

Moreover, the p-value was less than 0.05 for overdsipersion test on model 2, and this suggests that variance is much higher than the mean. As a poisson model's assumption of mean=variance is violated, it is important that we note how this would undermine the validity of GAM's inference reliability.

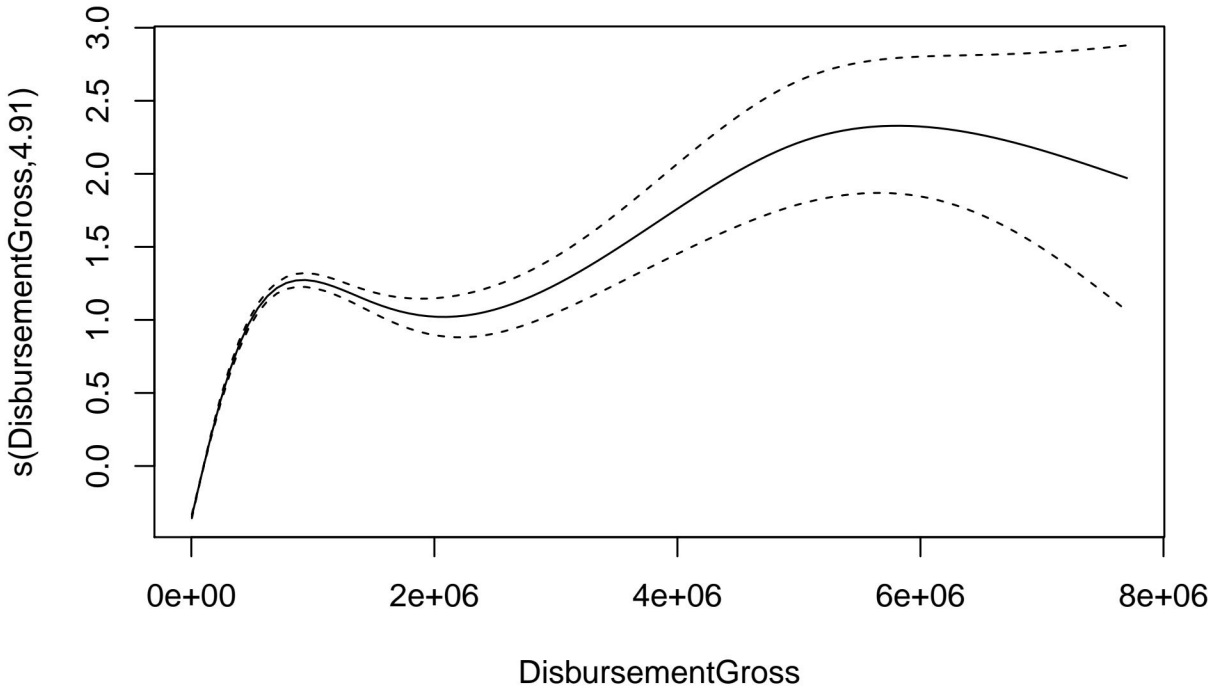


Figure 5 Plot of model 2's partial response function

From a plot of model 2's partial response function, we see that disbursementgross has a highly wiggly shape, suggesting the relationship with disbursementgross is highly non-linear. To ensure whether the non-linear term is really necesary, I used anova test with chi-square test statistic. Because I allowed mgcv to choose the smoothing parameter, I performed a bootstrap of 100 iterations, each with new dataframe replica with its # of jobs created replaced by n independent poisson random variables with mean being fitted values of model 1.

The resulting bootstraped p-value was 0. As it is less than 0.05, we reject the null hypothesis and claim that model 2 with its non-linear term is the correct model. **1**

In order to construct confidence intervals for the coefficients of interest from this model while accounting for the choice of smoothing parameter, we used non-parametric bootstrap methods. The result is shown later on in the result section.**2**

3 Modeling & Diagnostics 27 / 30

Model 1: Poisson regression on `CreateJob`

- **0 pts** Correct
- **2 pts** Misspecify the distribution of the response variable.

✓ - **1 pts** **Not enough explanation for the response distribution.**

- **2 pts** Missing explanation for substantial covariates
- **2 pts** Inappropriate/Missing model description
- **1 pts** Minor errors in model description

Model 2: Poisson Additive Model

✓ - **0 pts** **Correct**

- **2 pts** Misspecify the distribution of the response variable.
- **2 pts** Missing explanation for substantial covariates
- **2 pts** Inappropriate/Missing model description
- **1 pts** Minor errors in model description

Model Diagnostics

- **0 pts** Good enough
- **2 pts** **Model 1:** Missing diagnostics for the response distribution (Poisson). At least a discussion on residuals/responses of fit is required.
- **1 pts** $$ \llcorner $$ **Model 1:** Appropriate plots but with incorrect diagnostics; Poisson regression does not impose the normality/constant variance assumption on errors.
- **2 pts** **Model 2:** Missing diagnostics for the response distribution (Poisson). At least a discussion on residuals/responses is required.
- **1 pts** $$ \llcorner $$ **Model 2:** Appropriate plots but with incorrect diagnostics; Poisson additive model does not impose the normality/constant variance assumption on errors.

✓ - **2 pts** ****Model 1&2:** Missing diagnostics for the model bias: whether a deterministic formula for $$g(\mathbb{E}(Y|X))$$ as a function of the covariates is reasonable. At least a discussion on either of the followings is required: partial residuals/response plot, fitted values vs. true responses, good-ness of fit**

- **1 pts** Other minor errors

Hypothesis Testing via bootstrap

✓ - **0 pts** **Correct**

- **3 pts** Missing/Incorrect choice of Bootstrap method. Semi-/Non-parametric bootstrap is not appropriate here.
- **1 pts** Missing/Inappropriate justification for the bootstrap choice; Parametric bootstrap is justifiable since we are interested in the distribution of our statistics under the null hypothesis.
- **1 pts** **Parametric bootstrap**: Missing comment on what model to use for parameter estimation.
- **1 pts** Inappropriate test statistic
- **1 pts** Incorrect/Missing formulation of the p-value

Confidence Intervals for coefficients via bootstrap

✓ - **0 pts** **Correct**

- **3 pts** Missing justification for the bootstrap choice. Here, semi-/non-parametric bootstrap is good to use.
- **1 pts** Not enough explanation on CI construction

- **2 pts** Missing explanation on CI construction
- **1 pts** Other minor errors
- **6 pts** Missing

Results

From the test for non-linearity, the bootstrapped p-value of the anova hypothesis test between model 1 and 2 was 0, meaning that model 2, which uses non-linear term, is the correct model to choose. This goes in accordance with the non-linear relationship we see between the two models as shown in figure 3 earlier on.

Certainly, GLM is relatively easier to interpret; moreover, you can still get the rough idea of the general relationship variables. However, from the significance testing and the graph above, it is clear that fitted glm is not sufficient to capture the relationship. However, Therefore, I will use model 2 (GAM) for the rest of my analysis.

95% CI for the association between industry(shortdesc) and jobs created from model 2, while controlling for the other variables, is as shown below.

```
## [1] "Lower bound is"
```

```
##              shortdescAdminSupport                shortdescAgriculture
##                      -2.672237e-01                      -2.515698e+03
##            shortdescArtsRecreation               shortdescConstruction
##                      -5.924160e-01                      -8.879037e-01
##                 shortdescEducation           shortdescFinanceInsurance
##                      -1.061177e+00                      -1.165894e+00
##                   shortdescHealth                shortdescInformation
##                      -4.682292e-01                      -1.150662e+00
##             shortdescManufacturing                  shortdescMiningGas
##                      -6.755700e-01                      -7.519080e-01
##                    shortdescOther               shortdescProfServices
##                      -1.022477e+00                      -6.765420e-01
##               shortdescPublicAdmin                 shortdescRealEstate
##                      -4.234183e+05                      -8.865158e-01
##               shortdescRetailTrade shortdescTransportationWarehousing
##                      -9.426482e-01                      -9.007485e-01
##                 shortdescUtilities              shortdescWholesaleTrade
##                      -6.971754e+04                      -1.095579e+00
```

```
## [1] "Upper bound is "
```

```
##              shortdescAdminSupport                shortdescAgriculture
##                      -2.694525e-01                       -2.515717e+03
##            shortdescArtsRecreation               shortdescConstruction
##                      -5.953092e-01                       -8.951251e-01
##                 shortdescEducation           shortdescFinanceInsurance
##                      -1.068952e+00                       -1.176566e+00
##                   shortdescHealth                shortdescInformation
##                      -4.737351e-01                       -1.158457e+00
##             shortdescManufacturing                  shortdescMiningGas
##                      -6.808940e-01                       -7.568391e-01
##                   shortdescOther               shortdescProfServices
##                      -1.030696e+00                       -6.810961e-01
##               shortdescPublicAdmin                 shortdescRealEstate
##                      -4.167802e+05                       -8.892601e-01
##             shortdescRetailTrade shortdescTransportationWarehousing
##                      -9.503187e-01                       -9.079853e-01
##               shortdescUtilities             shortdescWholesaleTrade
##                      -6.971755e+04                       -1.103672e+00
```

These confidence intervals represent how much the number of jobs created increases depending on business's industry with 95% confidence. It seems Public admin has highest rate, whereas construction has the lowest rate. **2**

Conclusions

From the main findings in the analysis, our client can draw the following conclusions: -The relationship between jobs created and dollars loaned is non-linear. A loan that's twice as big will likely support fewer than twice as many jobs. **1** -Businesses of public admin industry create the most jobs per dollar loaned, whereas businesses of construction industry create the least. **2**

However, the client should be cautious when relying upon these findings to choose loans to grant due to several reasons. Firstly, in this analysis project, there were several available variables, which were unaccounted for in model diagnostic, that could have had significant impact on the findings. For example, default, Term, NoEmp and RetainedJob

4 Results 20 / 20

Results for test for nonlinearity

✓ **- 0 pts** **Correct**

- 3 pts The $$p$$-value is not close to zero (as it should be).

- 3 pts No plot demonstrates the shape of the relationship between loan amount and job creation.

- 10 pts Missing

Confidence intervals

✓ **- 0 pts** **Correct**

- 5 pts No table presents CIs for association between industry and jobs created.

- 3 pts Insufficient commentary on the CIs (e.g. no statement on what the CIs mean and no mention of which industries have the highest and lowest rates).

- 5 pts No comments on the CIs.

- 10 pts Missing

```
##              shortdescAdminSupport                shortdescAgriculture
##                      -2.694525e-01                       -2.515717e+03
##            shortdescArtsRecreation               shortdescConstruction
##                      -5.953092e-01                       -8.951251e-01
##                 shortdescEducation           shortdescFinanceInsurance
##                      -1.068952e+00                       -1.176566e+00
##                   shortdescHealth                shortdescInformation
##                      -4.737351e-01                       -1.158457e+00
##             shortdescManufacturing                  shortdescMiningGas
##                      -6.808940e-01                       -7.568391e-01
##                   shortdescOther               shortdescProfServices
##                      -1.030696e+00                       -6.810961e-01
##               shortdescPublicAdmin                 shortdescRealEstate
##                      -4.167802e+05                       -8.892601e-01
##             shortdescRetailTrade shortdescTransportationWarehousing
##                      -9.503187e-01                       -9.079853e-01
##               shortdescUtilities             shortdescWholesaleTrade
##                      -6.971755e+04                       -1.103672e+00
```

These confidence intervals represent how much the number of jobs created increases depending on business's industry with 95% confidence. It seems Public admin has highest rate, whereas construction has the lowest rate. **2**

Conclusions

From the main findings in the analysis, our client can draw the following conclusions: -The relationship between jobs created and dollars loaned is non-linear. A loan that's twice as big will likely support fewer than twice as many jobs. **1** -Businesses of public admin industry create the most jobs per dollar loaned, whereas businesses of construction industry create the least. **2**

However, the client should be cautious when relying upon these findings to choose loans to grant due to several reasons. Firstly, in this analysis project, there were several available variables, which were unaccounted for in model diagnostic, that could have had significant impact on the findings. For example, default, Term, NoEmp and RetainedJob

variables. It would be worth exploring these confounding variables. Secondly, considering the small sizes of these companies and the total population in US, the dataset's size is small in the grand scheme of things. More importantly, the data points are all based on businesses in PA; it is crucial for client to get data on companies outside PA, if they wish to generalizable findings. Lastly, all the data points are based on businesses between 1995 and 2014. Considering 8 years have passed since then, we would recommend our client to collect more recent data on companies, instead of immediately proceeding onward with our findings. Lastly, from the model diagnostics, we saw that our poisson GAM violates the homoscedasticity and mean=variance assumptions, and our client should note how these assumption violations undermine our model's ability to perform accurate inference before applying it.

5 Conclusions 15 / 15

✓ **- 0 pts Excellent! (Style, Findings and Discussion)**

Overall Requirements: Writing & Style

- **- 1 pts** Too much stats jargon.
- **- 3 pts** Is hard to understand for non-statisticians

Conclusions: Main Findings

- **- 0 pts** Perfect.
- **- 2 pts** No discussion on different industries.
- **- 2 pts** No discussion on (non-)linearity
- **- 3 pts** Partially missing
- **- 5 pts** Missing

Conclusions: Discussion & Limitations

- **- 0 pts** Excelent.
- **- 2 pts** You should mention that this was not a causal analysis.
- **- 2 pts** not enough discussion on limitations
- **- 5 pts** Missing

- **- 1 pts** Do not write out a list!
- **- 1 pts** No labels for answers
- **- 15 pts** Missing