Scatterplots– in class

In Class Work: Scatterplots

Complete the following steps using the `cex.Rdata` file:

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.0.5
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.4     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   2.0.1     v forcats 0.5.1
```

```
## Warning: package 'ggplot2' was built under R version 4.0.5
```

```
## Warning: package 'tibble' was built under R version 4.0.5
```

```
## Warning: package 'tidyr' was built under R version 4.0.5
```

```
## Warning: package 'readr' was built under R version 4.0.5
```

```
## Warning: package 'purrr' was built under R version 4.0.5
```

```
## Warning: package 'dplyr' was built under R version 4.0.5
```

```
## Warning: package 'stringr' was built under R version 4.0.5
```

```
## Warning: package 'forcats' was built under R version 4.0.5
```

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

```
library(forcats)
library(modelr)
```

```
## Warning: package 'modelr' was built under R version 4.0.5
```

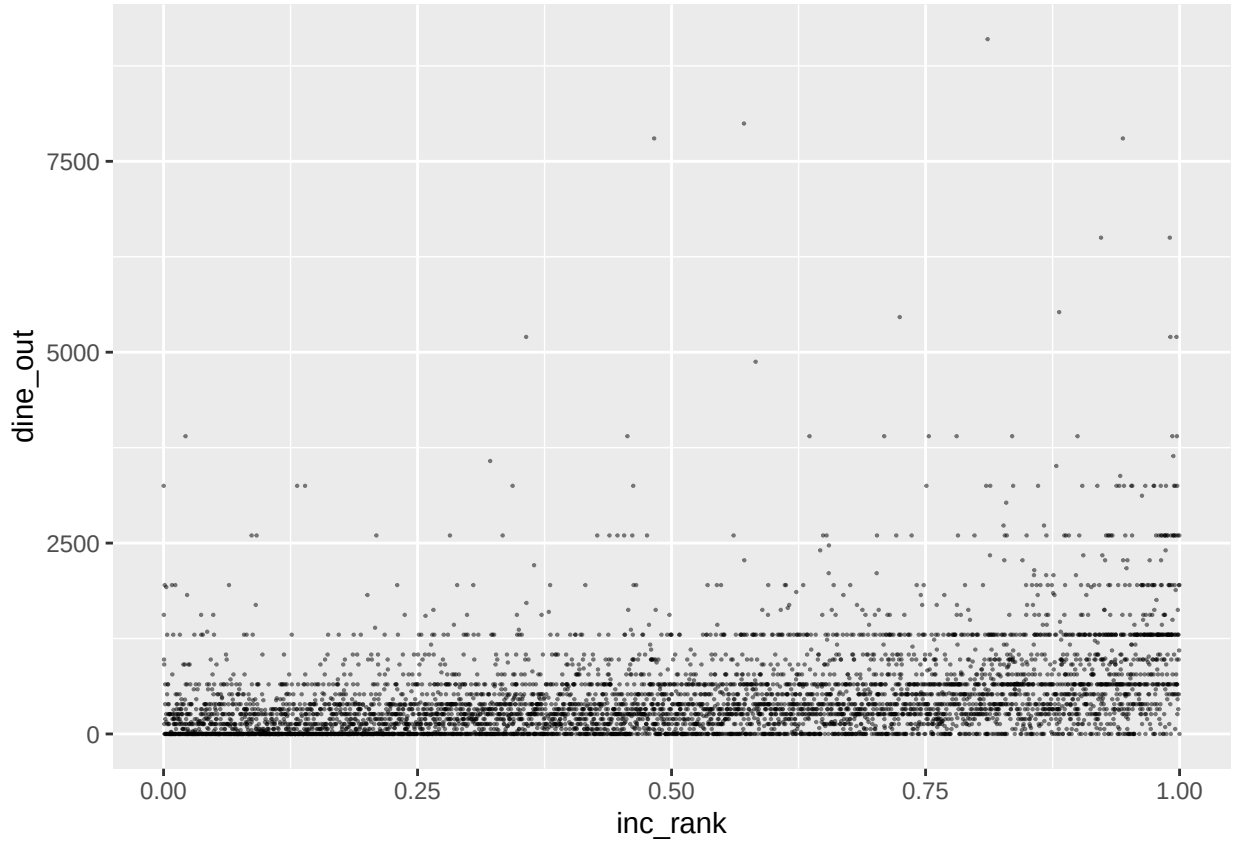
```
load("cex.RData")

names(cex)
```

```
##  [1] "newid"           "educ_ref"        "educa2"          "bls_urbn"
##  [5] "ref_race"        "race2"           "inclass"         "inc_rank"
##  [9] "sex_ref"         "sex2"            "hisp_ref"        "hisp2"
## [13] "pov_cym"         "region"          "fam_size"        "fam_type"
## [17] "childage"        "qyear"           "dine_out"        "grocery"
## [21] "grocery_nonfood" "grocery_food"    "booze_home"      "booze_out"
## [25] "other_store"     "cigarettes"      "trans_work"
```

1. Plot dining out as a function of income percentile rank.

```
cex%>%
  ggplot(aes(x=inc_rank,y=dine_out))+
  geom_point(size=.1,alpha=.5)
```

```
## Warning: Removed 1134 rows containing missing values (geom_point).
```



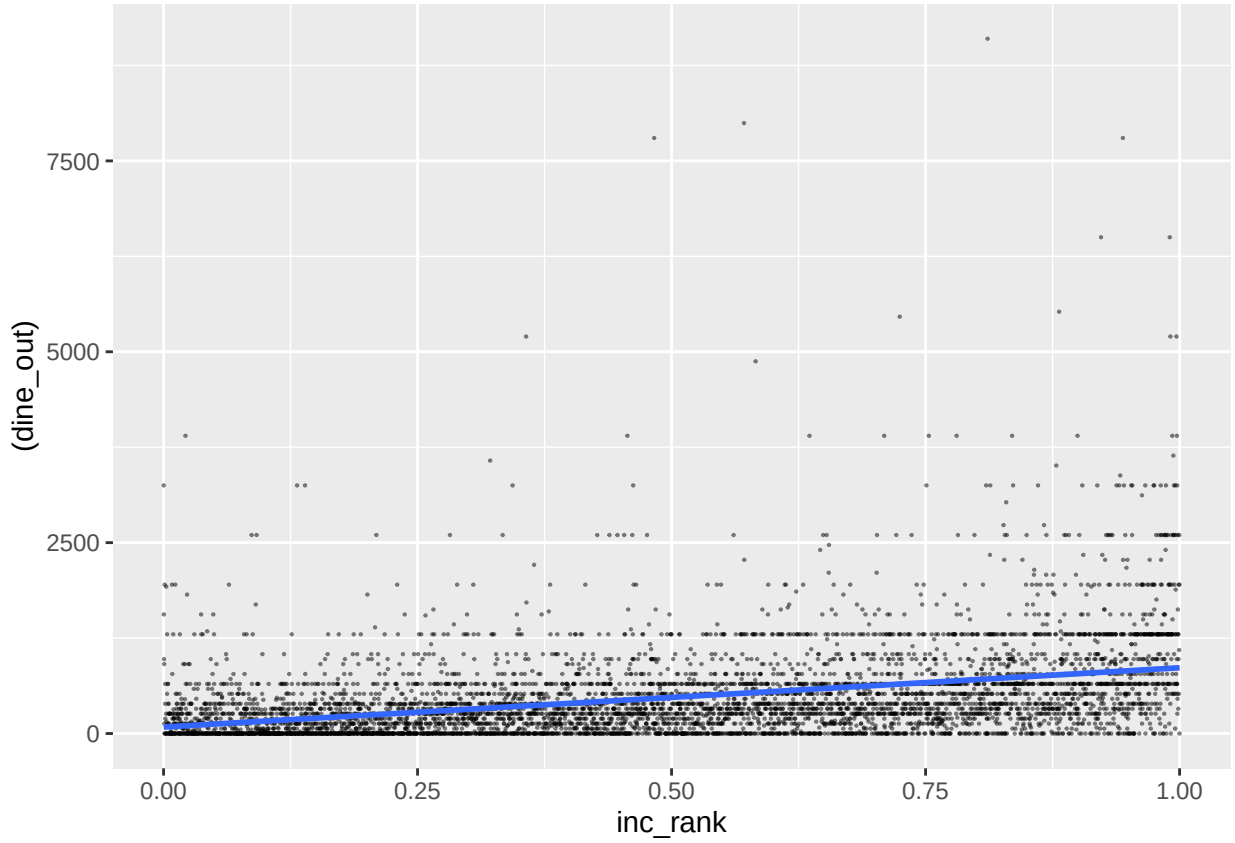
2. Add a line of best fit to the above graphic.

```
cex%>%
  ggplot(aes(x=inc_rank,y=(dine_out)))+
  geom_point(size=.1,alpha=.5) +
  geom_smooth(method="lm")
```

```
## `geom_smooth()` using formula 'y ~ x'
```

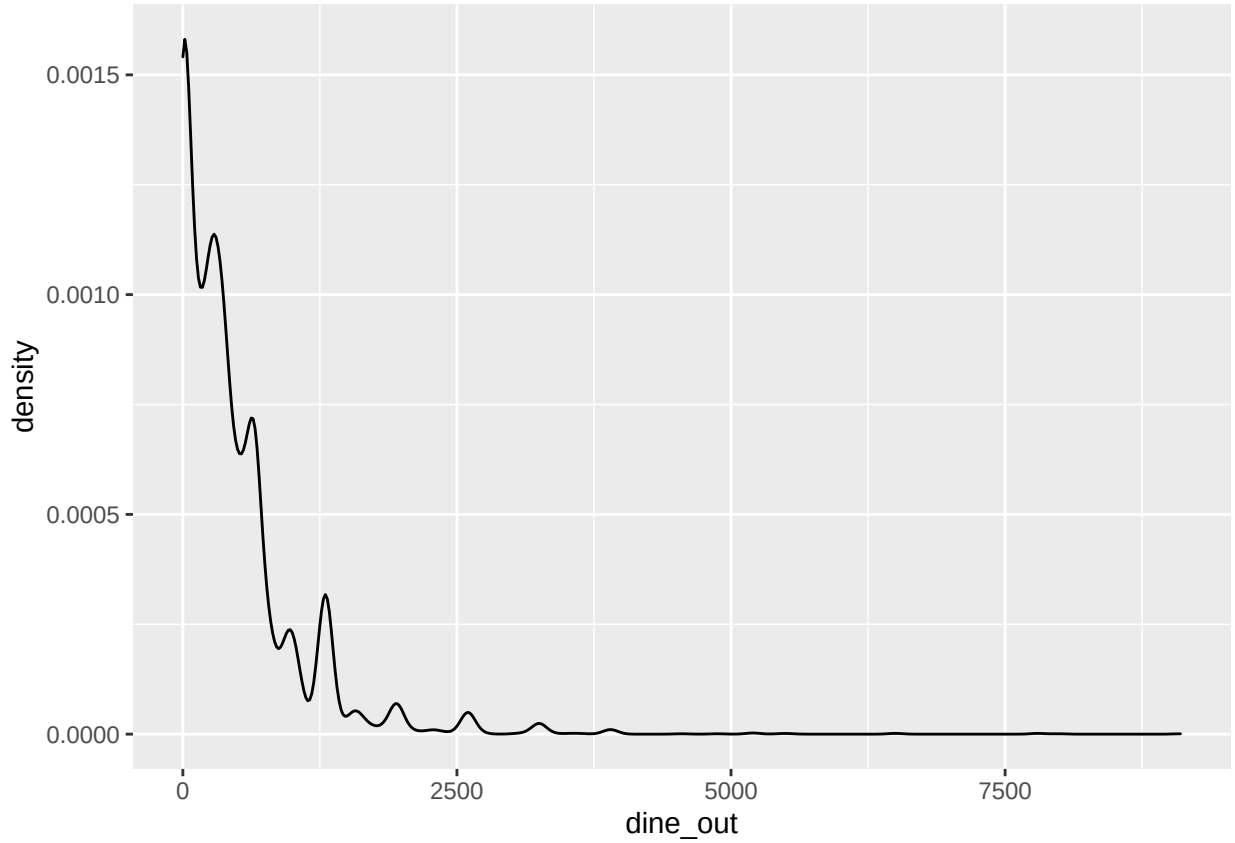
```
## Warning: Removed 1134 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 1134 rows containing missing values (geom_point).
```



3. Now create a plot with a scale appropriate to dining out as the dependent variable.

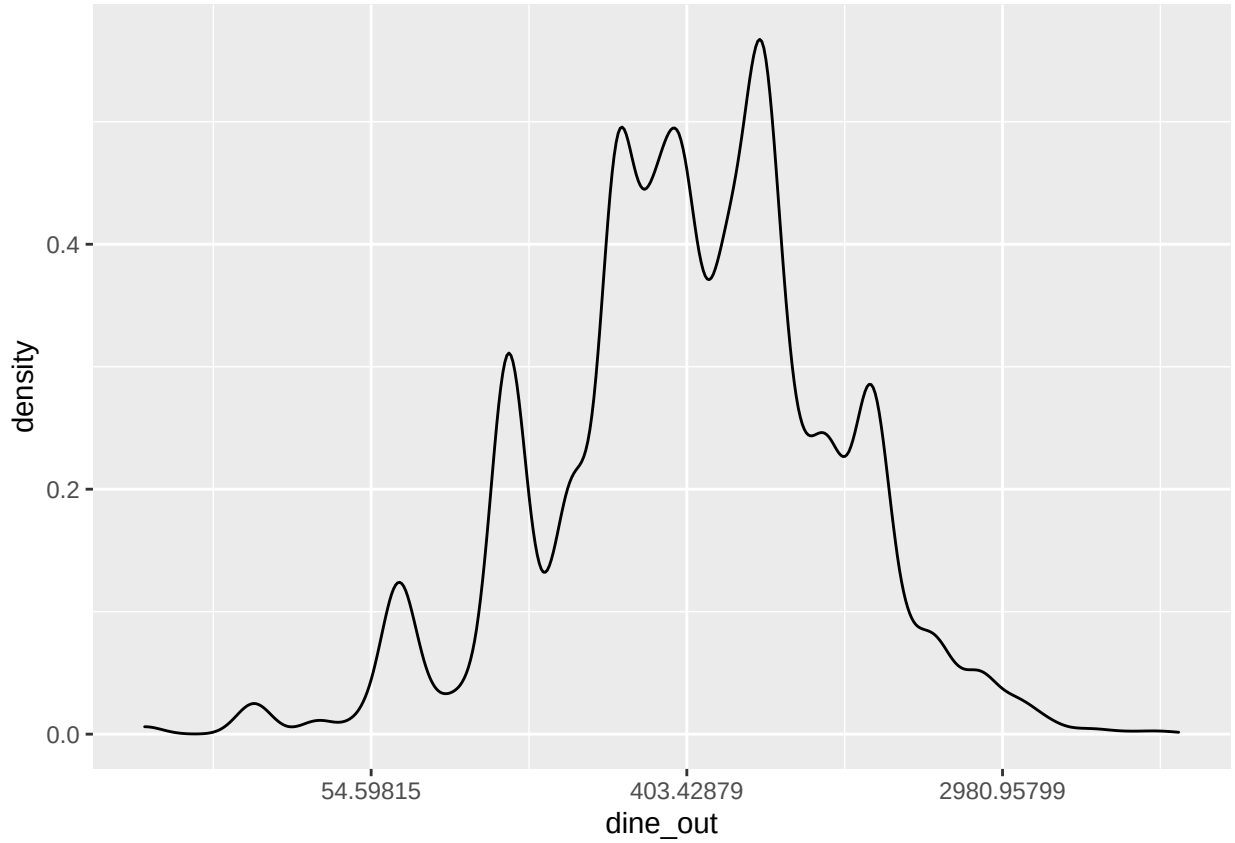
```
cex%>%
  ggplot(aes(x=dine_out))+
  geom_density()
```




```
cex%>%
  ggplot(aes(x=dine_out))+
  geom_density()+
  scale_x_continuous(trans="log")
```

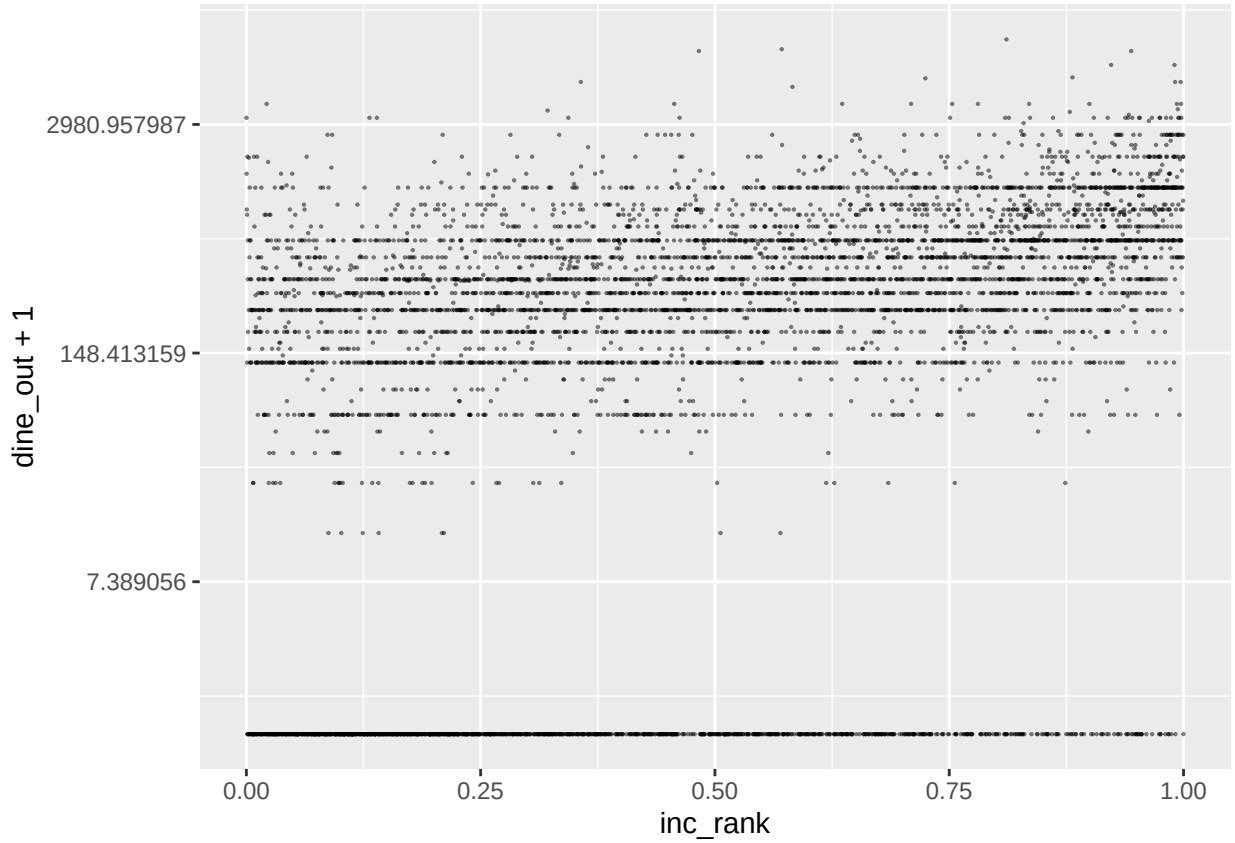
```
## Warning: Transformation introduced infinite values in continuous x-axis
```

```
## Warning: Removed 1509 rows containing non-finite values (stat_density).
```




```
cex%>%
  ggplot(aes(x=inc_rank,y=dine_out+1))+
  geom_point(size=.1,alpha=.5)+
  scale_y_continuous(trans="log")
```

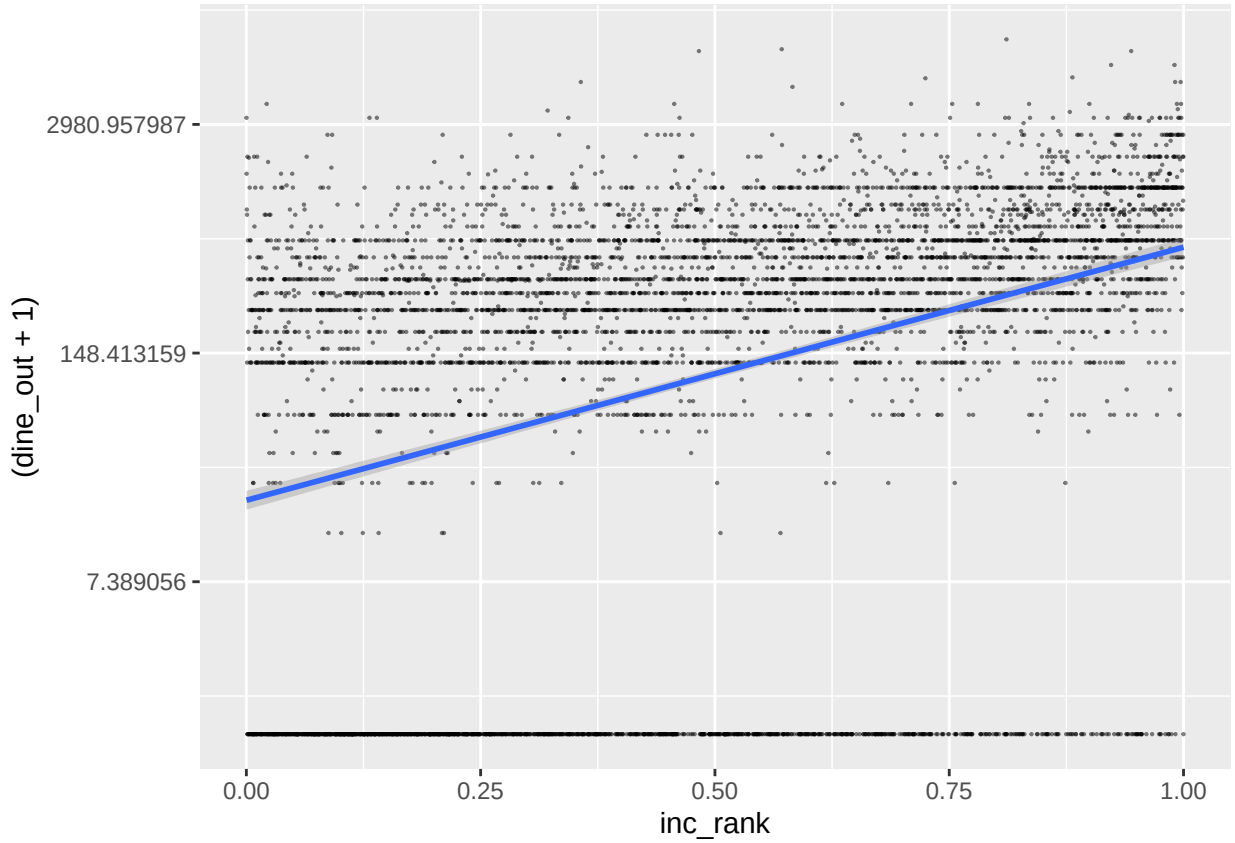
```
## Warning: Removed 1134 rows containing missing values (geom_point).
```



4. Add a line of best fit to your new graphic.

```
cex%>%
  ggplot(aes(x=inc_rank,y=(dine_out+1)))+
  geom_point(size=.1,alpha=.5) +
  scale_y_continuous(trans="log")+
  geom_smooth(method="lm")
```

```
## `geom_smooth()` using formula 'y ~ x'

## Warning: Removed 1134 rows containing non-finite values (stat_smooth).

## Warning: Removed 1134 rows containing missing values (geom_point).
```



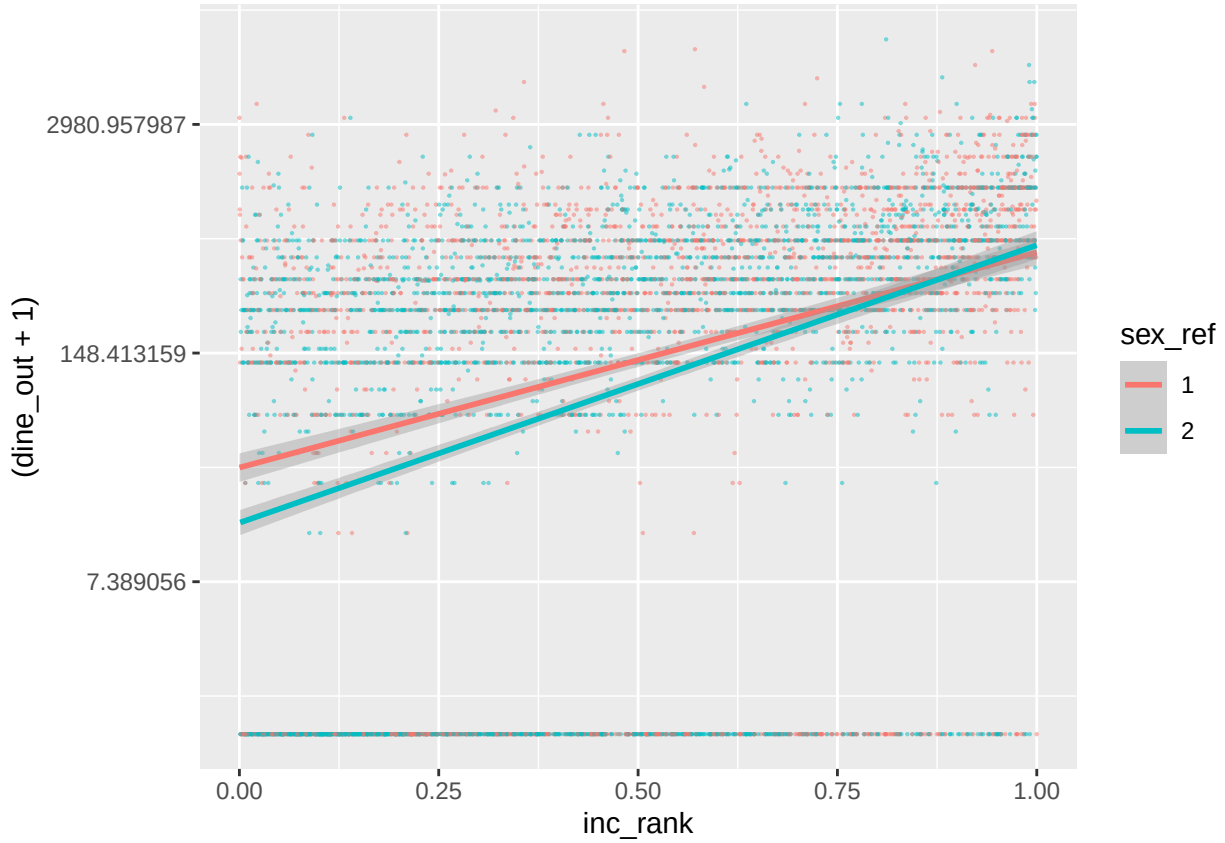
5. Create another plot with dining out on the y axis, income percentile rank on the x axis, and a categorical (factor) variable that differentiates the points.

```
cex%>%
  ggplot(aes(x=inc_rank,y=(dine_out+1),color=sex_ref ))+
  geom_point(size=.1,alpha=.5) +
  scale_y_continuous(trans="log")+
  geom_smooth(method="lm")
```

```
## `geom_smooth()` using formula 'y ~ x'
```

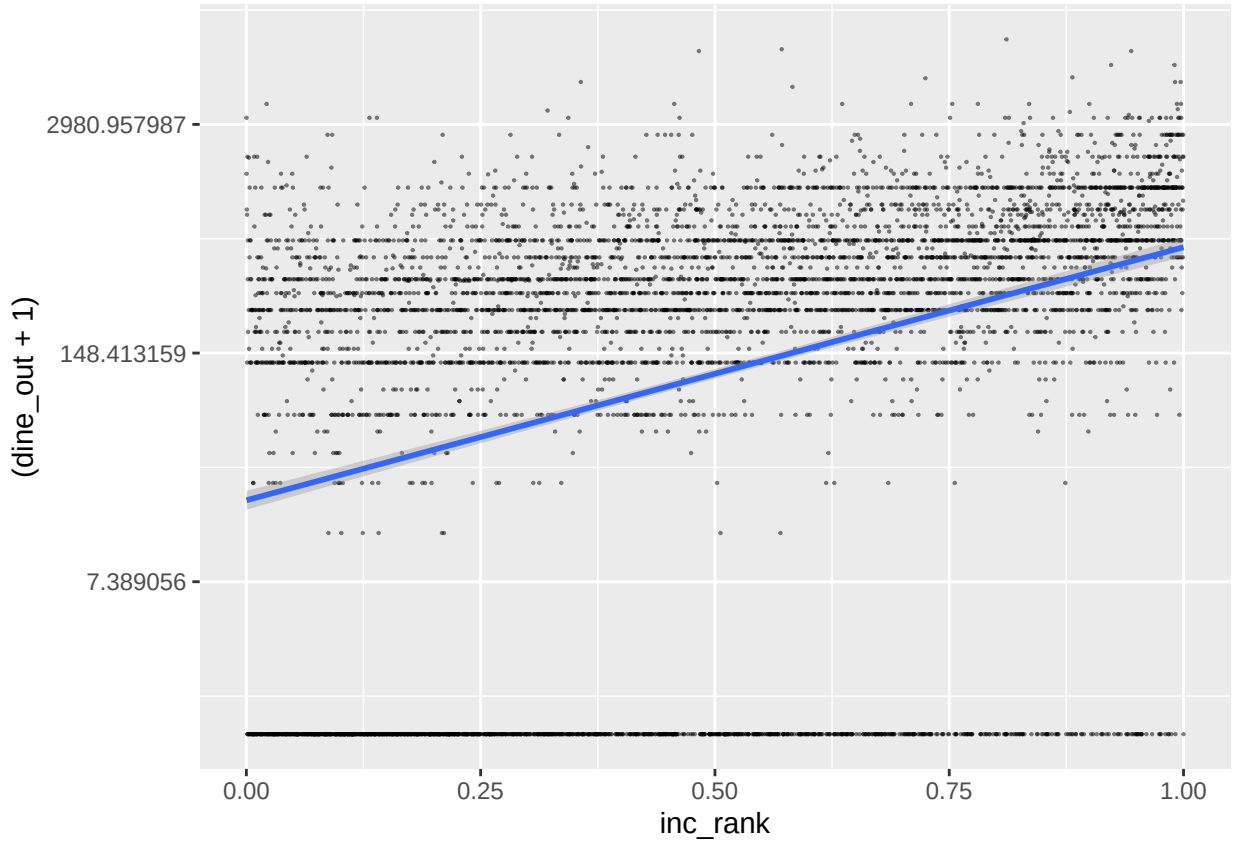
```
## Warning: Removed 1134 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 1134 rows containing missing values (geom_point).
```




```
cex%>%
  ggplot(aes(x=inc_rank,y=(dine_out+1)))+
  geom_point(size=.1,alpha=.5) +
  scale_y_continuous(trans="log")+
  geom_smooth(method="lm")
```

```
## `geom_smooth()` using formula 'y ~ x'

## Warning: Removed 1134 rows containing non-finite values (stat_smooth).

## Warning: Removed 1134 rows containing missing values (geom_point).
```



6. Run a regression that has (possibly transformed) dining out out as the dependent variable, with income and at least one other variable as predictors. What's the RMSE (log scale, if needed) from this model (relative to the testing dataset)?

```
library(tidymodels)
```

```
## Warning: package 'tidymodels' was built under R version 4.0.5
```

```
## Registered S3 method overwritten by 'tune':
##   method                    from
##   required_pkgs.model_spec parsnip
```

```
## -- Attaching packages ------------------------------------- tidymodels 0.1.3 --
```

```
## v broom        0.7.9      v rsample      0.1.0
## v dials        0.0.10     v tune         0.1.6
## v infer        1.0.0      v workflows    0.2.3
## v modeldata    0.1.1      v workflowsets 0.1.0
## v parsnip      0.1.7      v yardstick    0.0.8
## v recipes      0.1.17
```

```
## Warning: package 'broom' was built under R version 4.0.5
```

```
## Warning: package 'dials' was built under R version 4.0.5
```

```
## Warning: package 'scales' was built under R version 4.0.5

## Warning: package 'infer' was built under R version 4.0.5

## Warning: package 'modeldata' was built under R version 4.0.5

## Warning: package 'parsnip' was built under R version 4.0.5

## Warning: package 'rsample' was built under R version 4.0.5

## Warning: package 'tune' was built under R version 4.0.5

## Warning: package 'workflows' was built under R version 4.0.5

## Warning: package 'workflowsets' was built under R version 4.0.5

## Warning: package 'yardstick' was built under R version 4.0.5

## -- Conflicts --------------------------------------- tidymodels_conflicts() --
## x broom::bootstrap() masks modelr::bootstrap()
## x scales::discard()  masks purrr::discard()
## x dplyr::filter()    masks stats::filter()
## x recipes::fixed()   masks stringr::fixed()
## x dplyr::lag()       masks stats::lag()
## x yardstick::mae()   masks modelr::mae()
## x yardstick::mape()  masks modelr::mape()
## x yardstick::rmse()  masks modelr::rmse()
## x yardstick::spec()  masks readr::spec()
## x recipes::step()    masks stats::step()
## * Use tidymodels_prefer() to resolve common conflicts.
```

```
set.seed(35202)

#Split data
split_data<-cex%>%initial_split(prop=.5)

cex_train<-training(split_data)

cex_test<-testing(split_data)


#Specify model
lm_fit <-
  linear_reg() %>%
  set_engine("lm")%>%
  set_mode("regression")

#Specify formula
dine_formula<-as.formula("log(dine_out +1)~inc_rank + grocery")

#Fit to Training Data
```

```
lm_results<-
lm_fit%>%
fit(dine_formula,data=cex_train)

#Predict Testing Data
cex_test<-
    lm_results%>%
      predict(new_data=cex_test)%>%
      rename(pred1=.pred)%>%
      bind_cols(cex_test)

#Calculate Fit
rmse_1<-
rmse(cex_test, truth=log(dine_out +1), estimate=pred1)

rmse_1
```

```
## # A tibble: 1 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 rmse    standard        2.39
```

7. Create new predictions from your model.

```
mod1<-lm(dine_formula,data=cex)

summary(mod1)
```

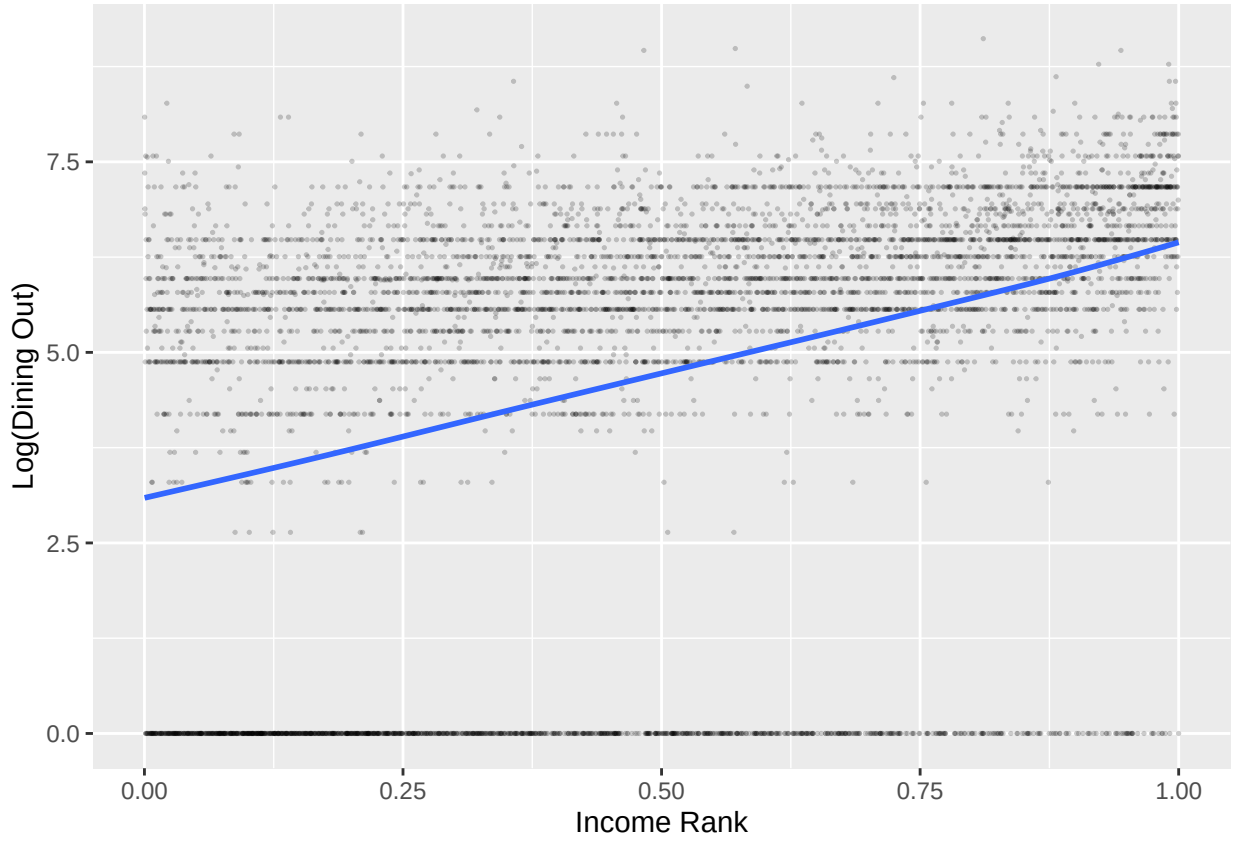
```
##
## Call:
## lm(formula = dine_formula, data = cex)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.4720 -0.4706  0.7940  1.6017  5.1572
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.9553101  0.0698901   42.28  < 2e-16 ***
## inc_rank    3.1566803  0.1174787   26.87  < 2e-16 ***
## grocery     0.0001354  0.0000361    3.75 0.000179 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.403 on 5701 degrees of freedom
##   (1134 observations deleted due to missingness)
## Multiple R-squared:  0.1416, Adjusted R-squared:  0.1413
## F-statistic: 470.4 on 2 and 5701 DF,  p-value: < 2.2e-16
```

```
cex<-cex%>%add_predictions(mod1)%>%rename(pred_mod1=pred)
```

8. Create a plot showing predicted levels of dining out based on income and your other variable from the model in number 6.

```
gg<-ggplot(cex,aes(x=inc_rank,y=log(dine_out + 1)))
gg<-gg+geom_point(alpha=.2,size=.25)

gg<-gg+geom_smooth(data=cex,(aes(x=inc_rank,y=pred_mod1)))

gg<-gg+xlab("Income Rank")+ylab("Log(Dining Out)")

gg
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'

## Warning: Removed 1134 rows containing non-finite values (stat_smooth).

## Warning: Removed 1134 rows containing missing values (geom_point).
```



9. Make your plot beautiful in every way. Make sure that axes are labeled appropriately, that colors are used well, and that legends help the reader to make sense of the plot.

```
gg <- gg + ggtitle("Dining Out vs. Income Rank")
gg
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```

```
## Warning: Removed 1134 rows containing non-finite values (stat_smooth).

## Warning: Removed 1134 rows containing missing values (geom_point).
```