Broadway data analysis

Julianna Szabo

12/23/2020

Executive summary

Research question

Is there a correlation between the occupancy percentage of a show and the revenue per attendant?
I will be looking for causality.

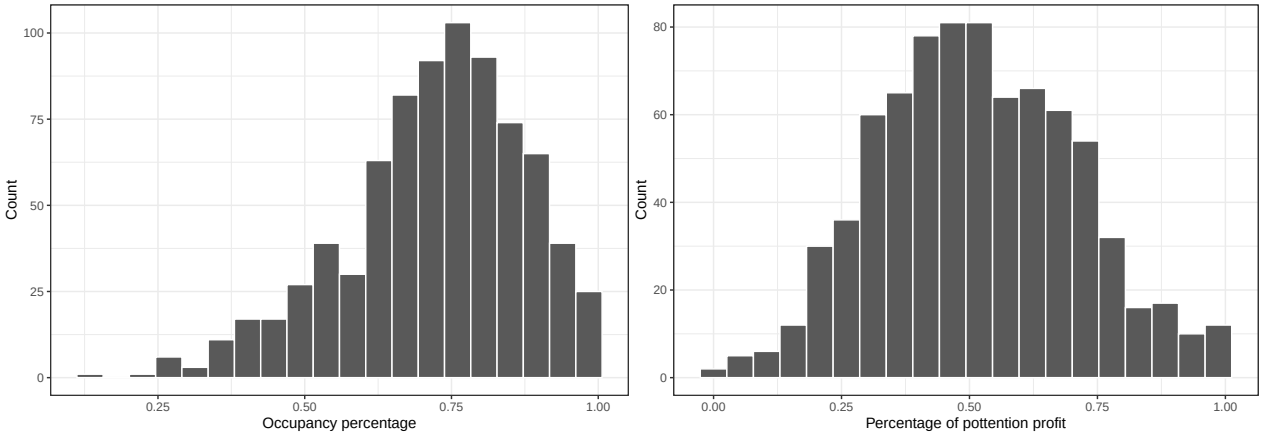
Data

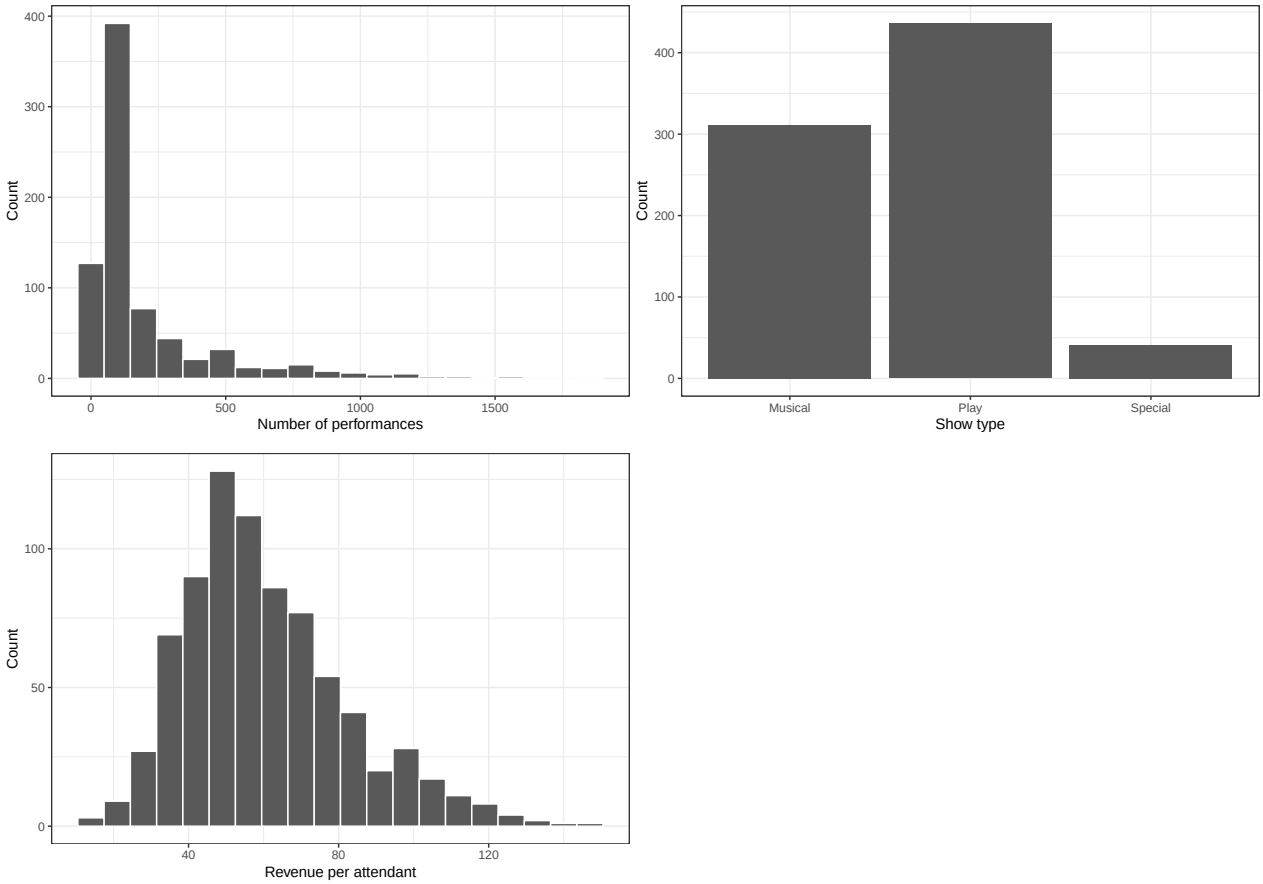
The data is very complete and representative. I have removed some missing values during the cleaning process but it was a very small percentage. Further some measures were lost by switching from a time series to a cross sectional data set. However, I aggregated on the show name, which lets me keel the most amount of detail. Most of the variables are quantitative so that means they measure what they decribe. I will use using Revenue / Attendant, where Revenue is measured as the gross revenue of the show, and attendants which are measured as total number of people who attended the show.

My x variable will be Occupancy percentage (capacity_filled) My y variable will be Revenue / Attendant which I will calculate based on revenue and attendant

There may be some measurement error in y, which is classic and doesn't affect the slope. There may be some measuement error in x which could also be classic, which does affect the slope.

Summary of variables





variable	type	n	mean	median	min	max	sd
Occupancy percentage	x	788	0.72	0.74	0.15	1.00	0.15
Percentage of possible profit	x	788	0.51	0.50	0.01	1.00	0.19
Number of performances	x	788	342.90	102.00	0.00	8400.00	934.22
Revenue per Attendant	y	788	60.60	56.42	12.64	145.64	21.86

Looks like they are distributed somehwat normall, but y has a long right tail, while x has more of a left tail. Also looking at x, there are a few outliers, since a percentage should not be larger than 1. Therefore I will remove these from the set.

Correlation

Ln transformations

Appendix

Log transformaton on y level of occupancy percentage level of percentage of possible profit log of number of performances

Regression Models

Decided on ..... model

Appendix

Correlation

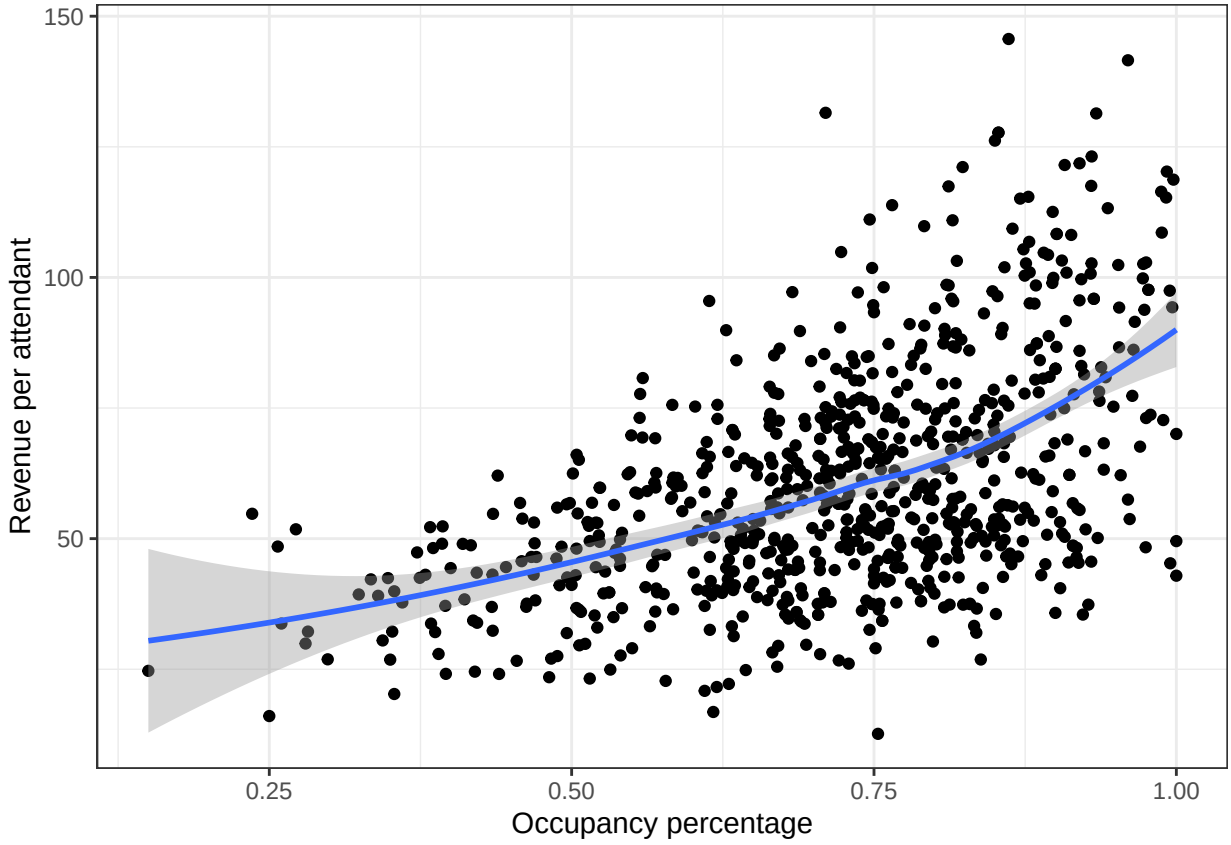
Var1	Var2	corr_val
ln_capacity_filled	capacity_filled	0.98
capacity_filled_sq	capacity_filled	0.99
ln_percentage_of_poss_profit	percentage_of_poss_profit	0.92
ln_revenue_per_att	revenue_per_att	0.97
capacity_filled	ln_capacity_filled	0.98
capacity_filled_sq	ln_capacity_filled	0.94
revenue_per_att	ln_revenue_per_att	0.97
percentage_of_poss_profit	ln_percentage_of_poss_profit	0.92
capacity_filled	capacity_filled_sq	0.99
ln_capacity_filled	capacity_filled_sq	0.94

Ln transformation

Occupancy percentage

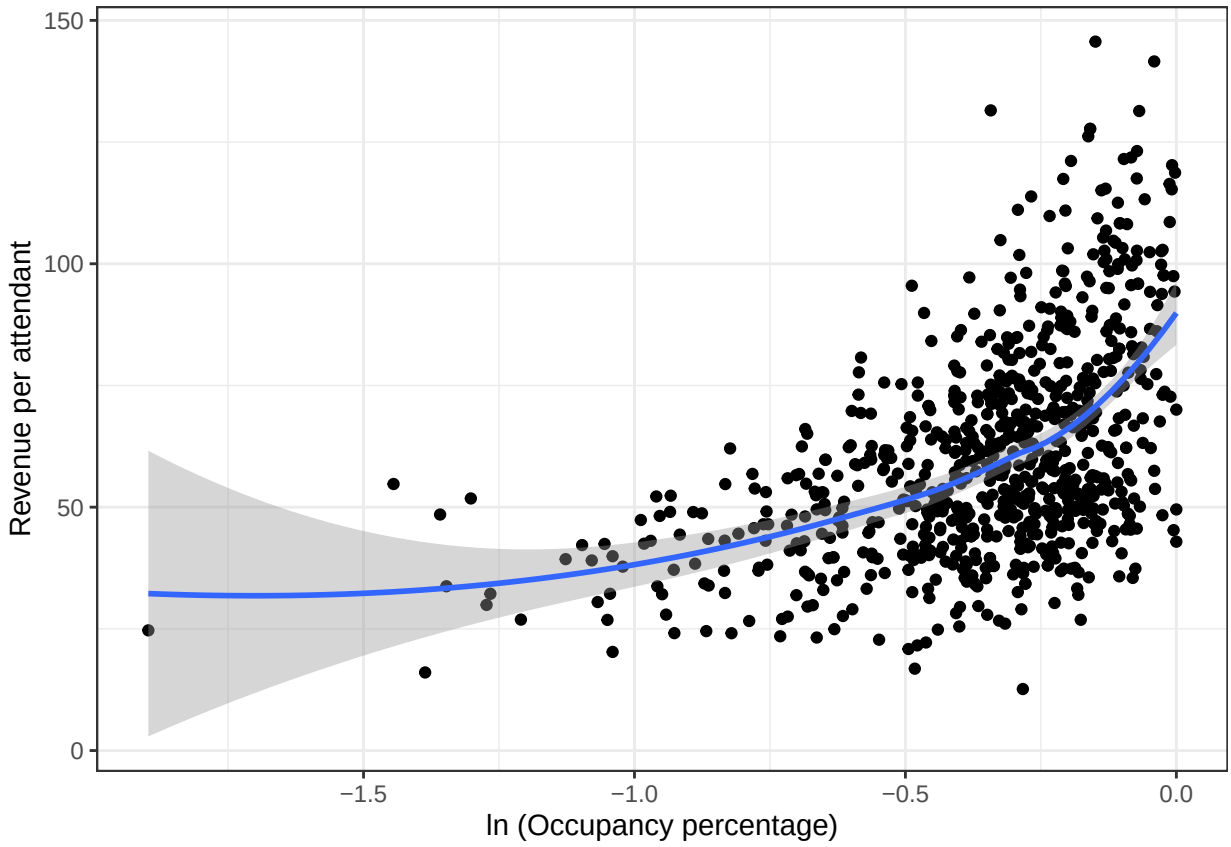
Level - level regression

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



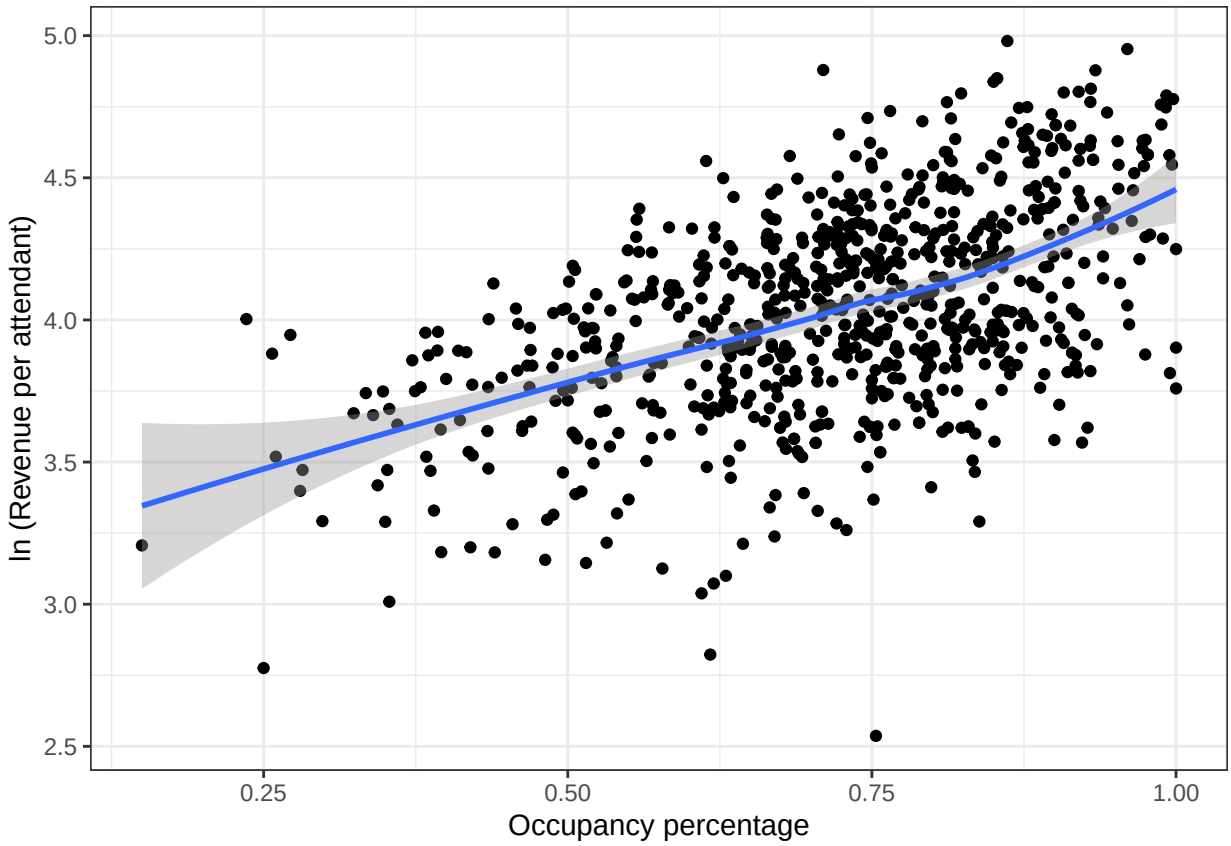
Log - level regression

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



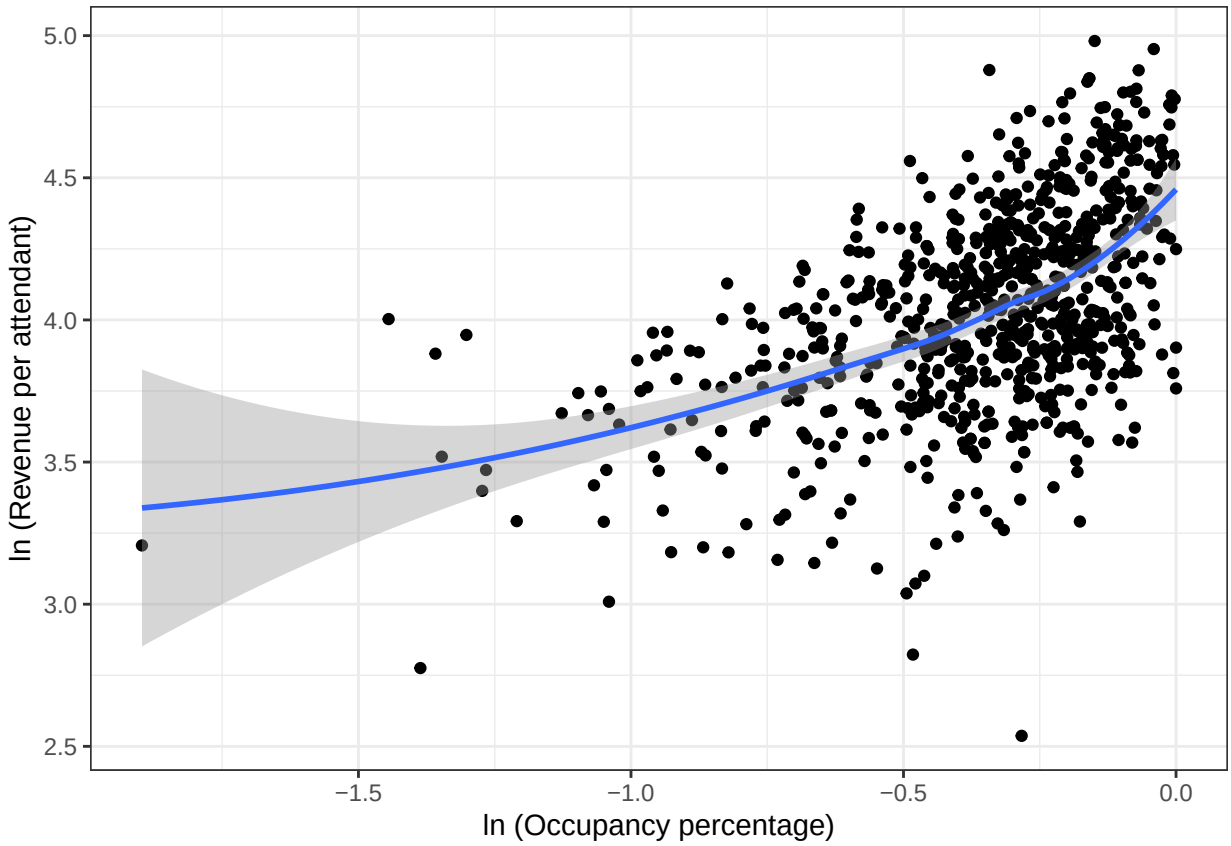
Level - log regression

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



Log - log regression

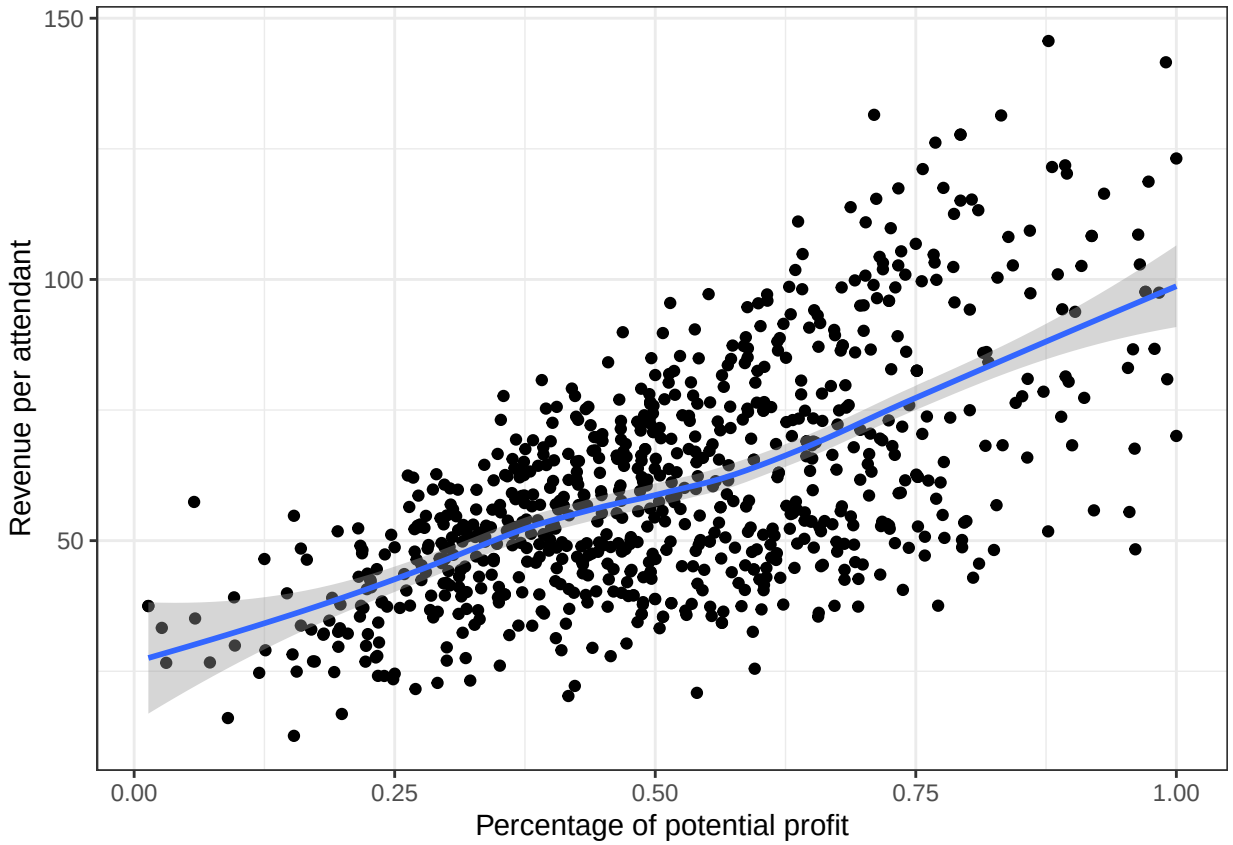
```
## `geom_smooth()` using formula 'y ~ x'
```



Percentage of potential profit

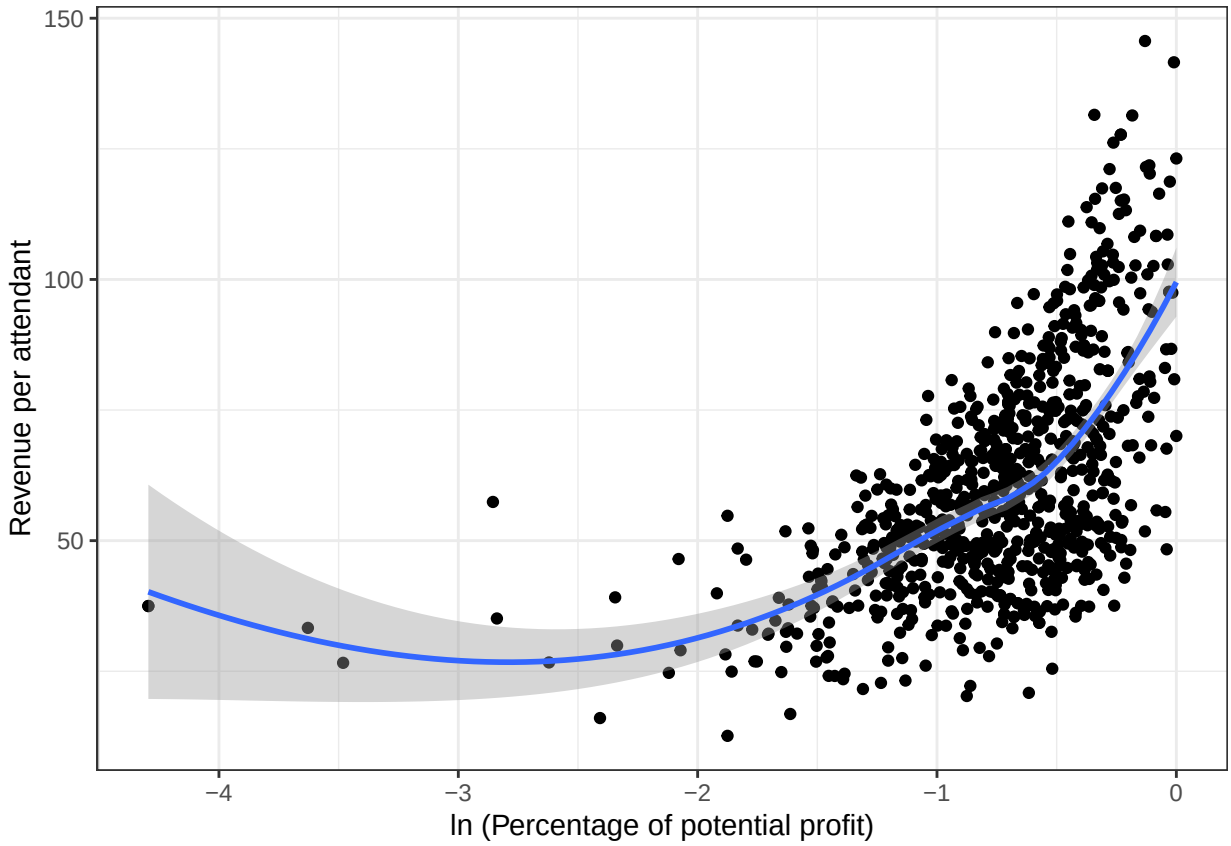
Level - level regression

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



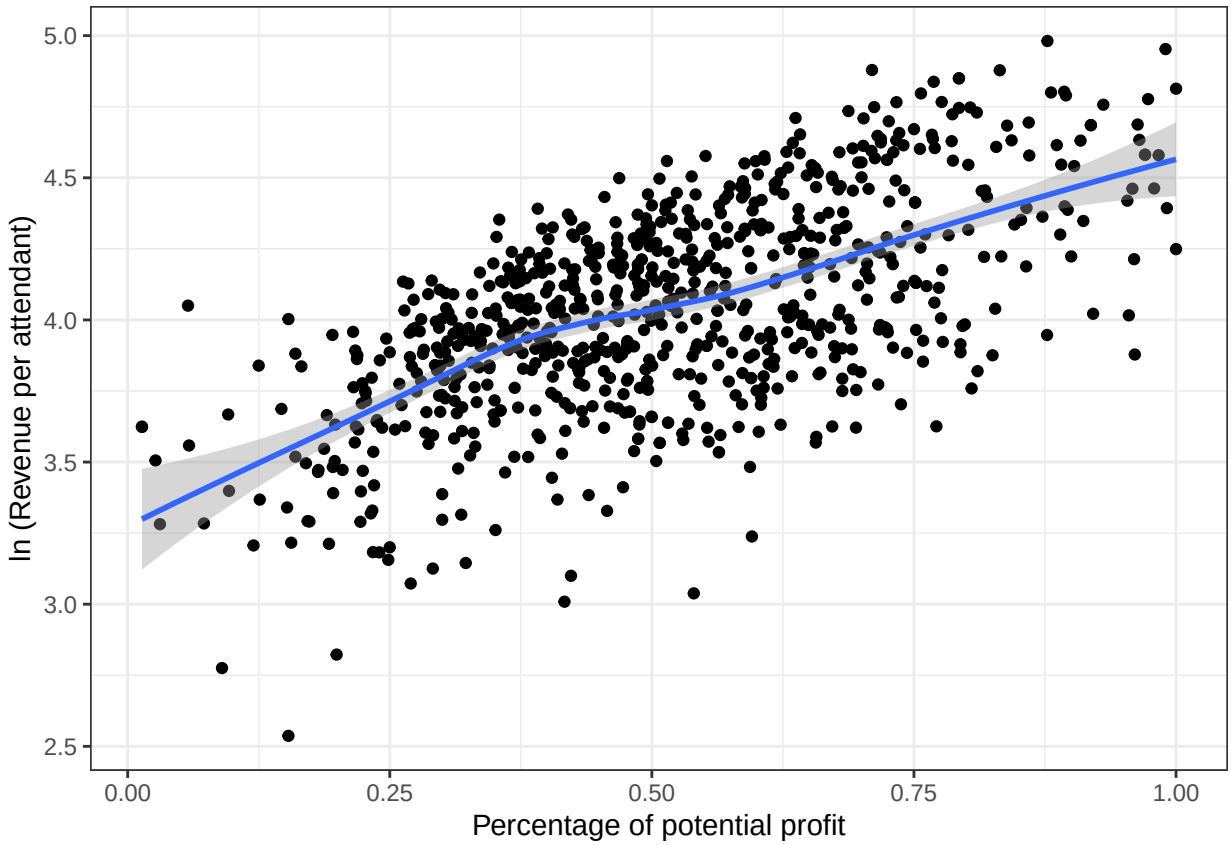
Log - level regression

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



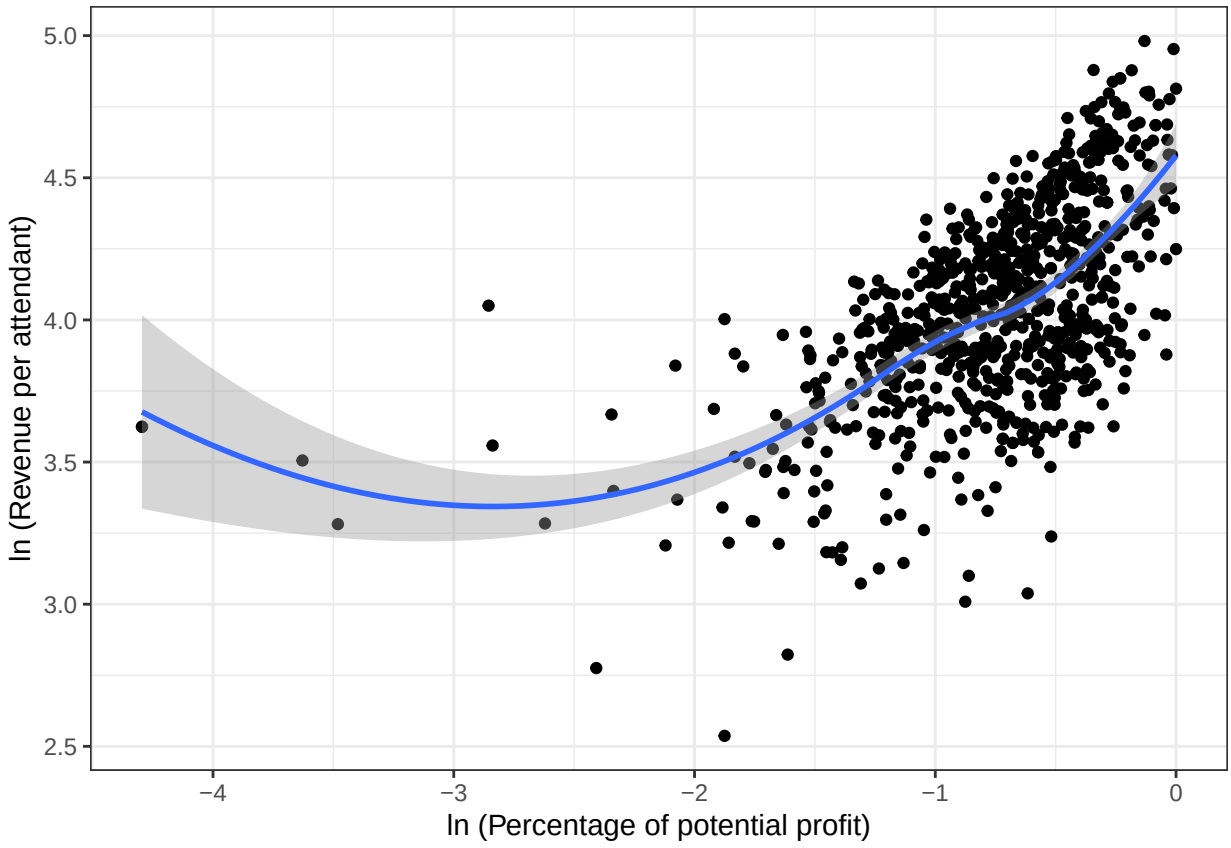
Level - log regression

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



Log - log regression

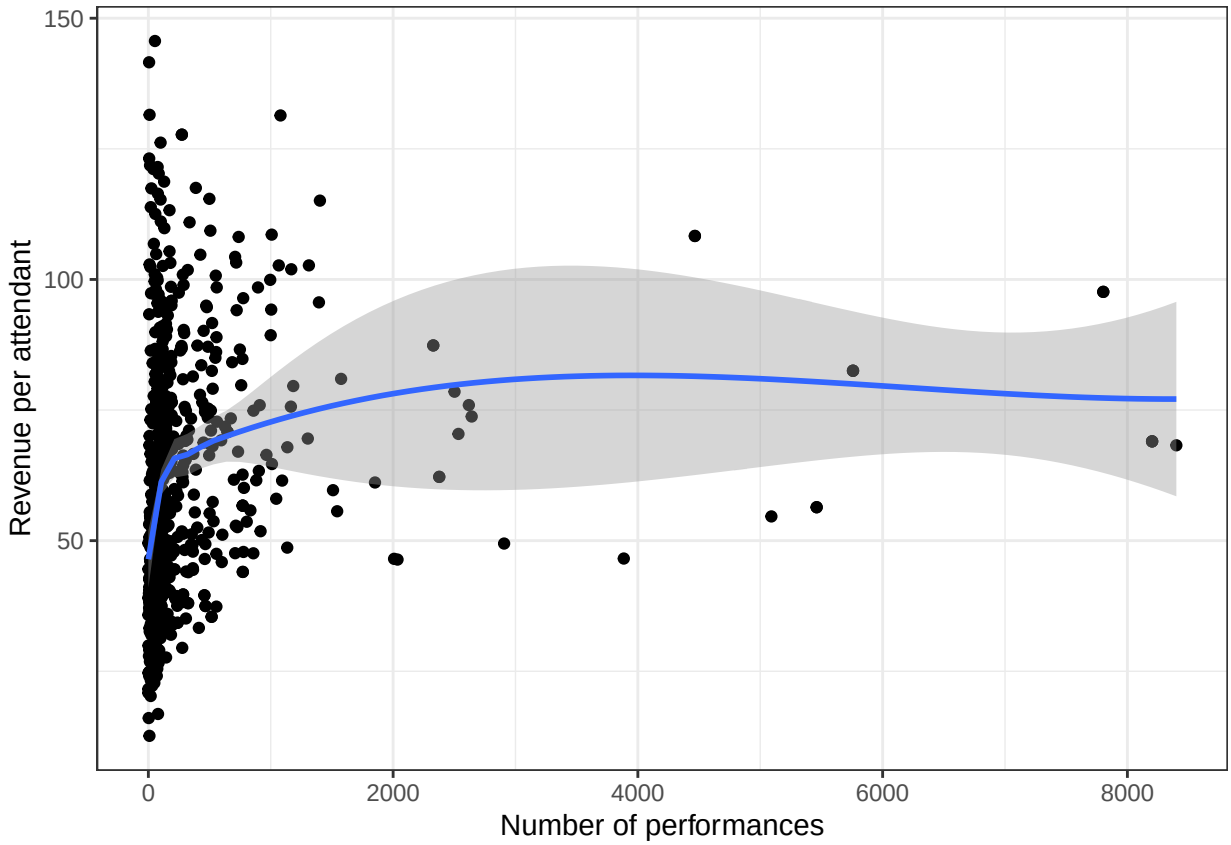
```
## `geom_smooth()` using formula 'y ~ x'
```



Number of performances

Level - level regression

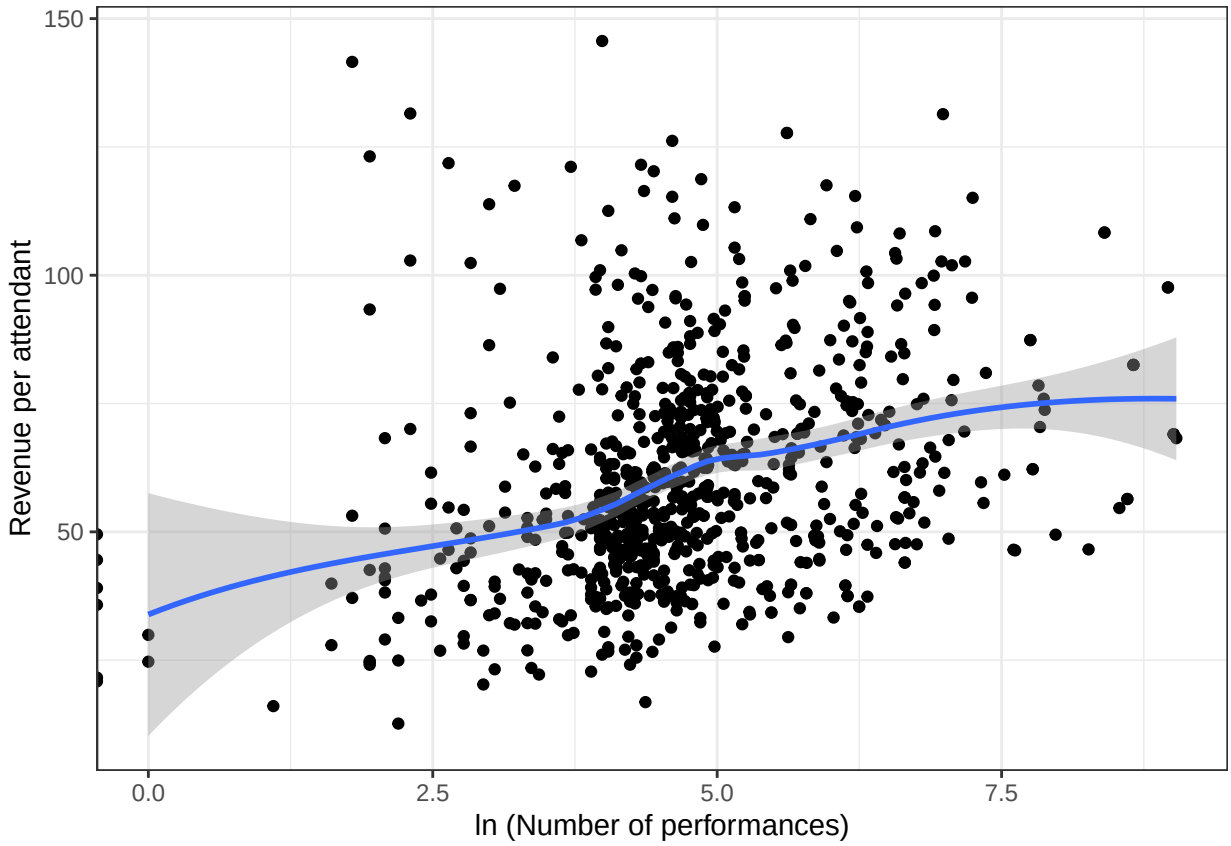
```
## `geom_smooth()` using formula 'y ~ x'
```



Log - level regression

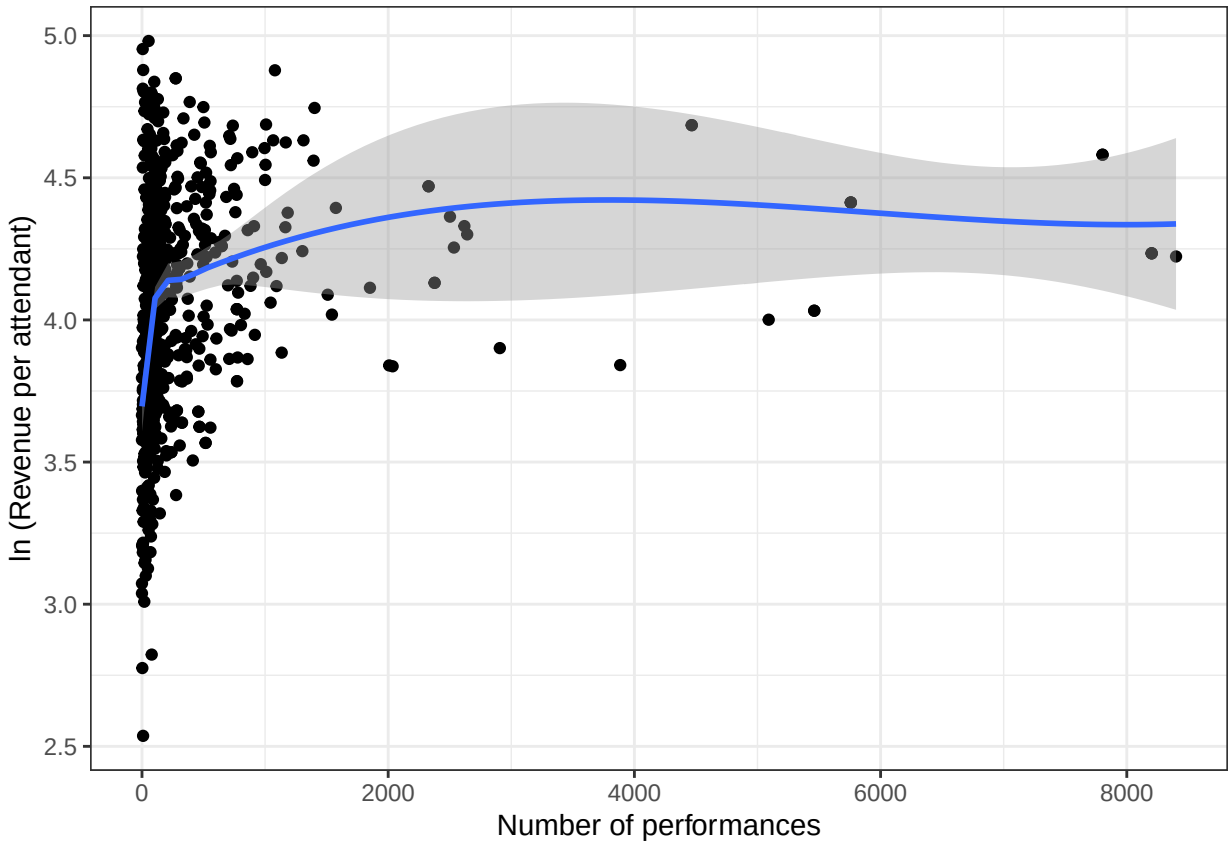
```
## `geom_smooth()` using formula 'y ~ x'
```

```
## Warning: Removed 6 rows containing non-finite values (stat_smooth).
```



Level - log regression

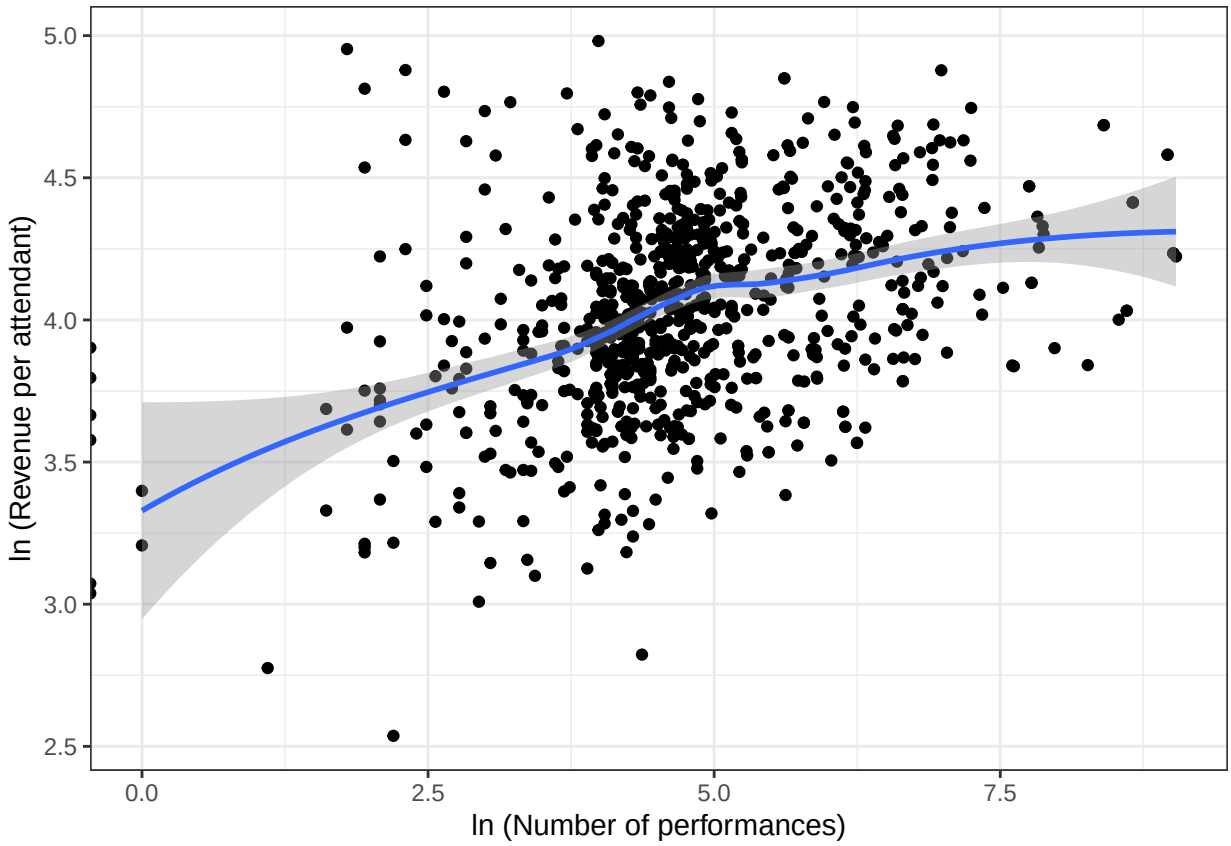
```
## `geom_smooth()` using formula 'y ~ x'
```



Log - log regression

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

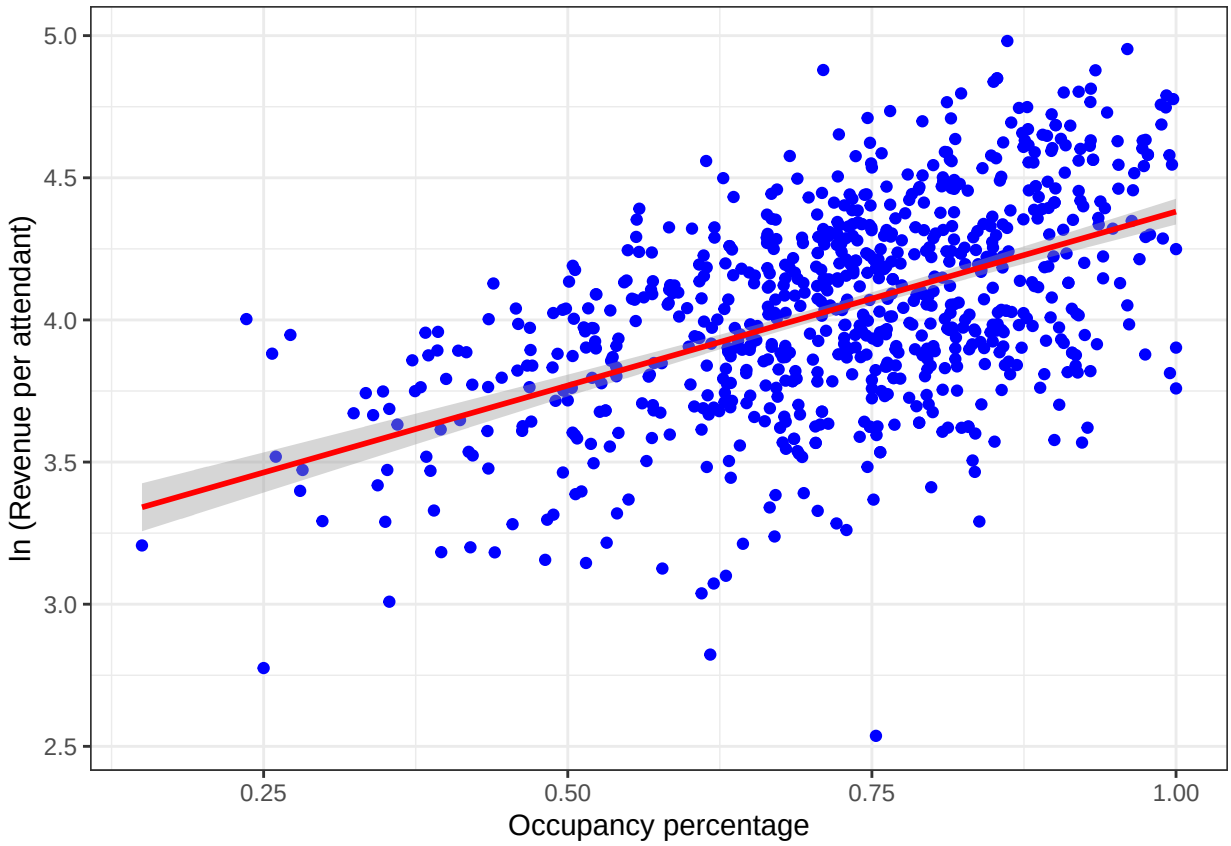
```
## Warning: Removed 6 rows containing non-finite values (stat_smooth).
```



Regression modes

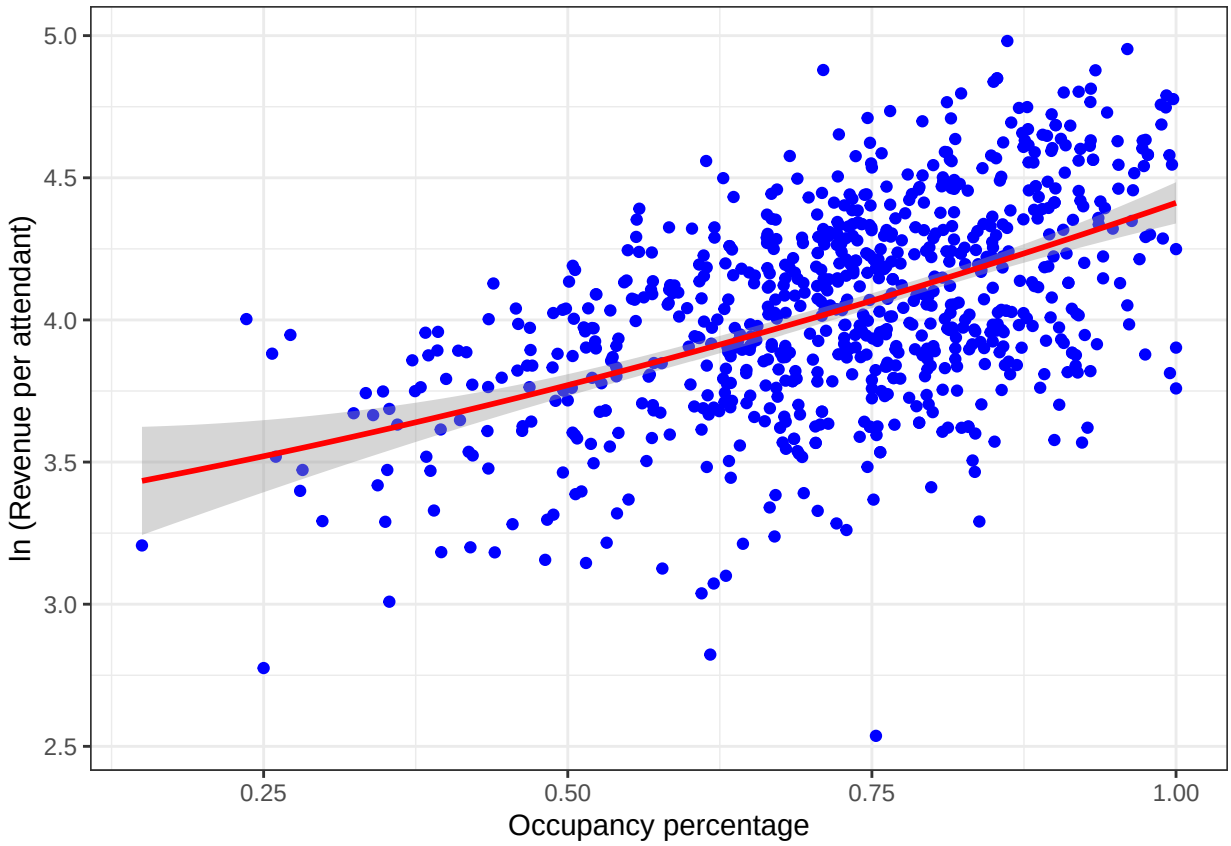
Regression 1 - Simple linear regression

```
##
## Call:
## lm_robust(formula = ln_revenue_per_att ~ capacity_filled, data = df,
##     se_type = "HC2")
##
## Standard error type:  HC2
##
## Coefficients:
##                 Estimate Std. Error t value   Pr(>|t|) CI Lower CI Upper  DF
## (Intercept)        3.157    0.05182   60.93 5.364e-300    3.056    3.259 786
## capacity_filled    1.223    0.07155   17.10  5.624e-56    1.083    1.364 786
##
## Multiple R-squared:  0.2653 ,    Adjusted R-squared:  0.2644
## F-statistic: 292.4 on 1 and 786 DF,  p-value: < 2.2e-16

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



Regression 2 - Quadratic (linear) regression

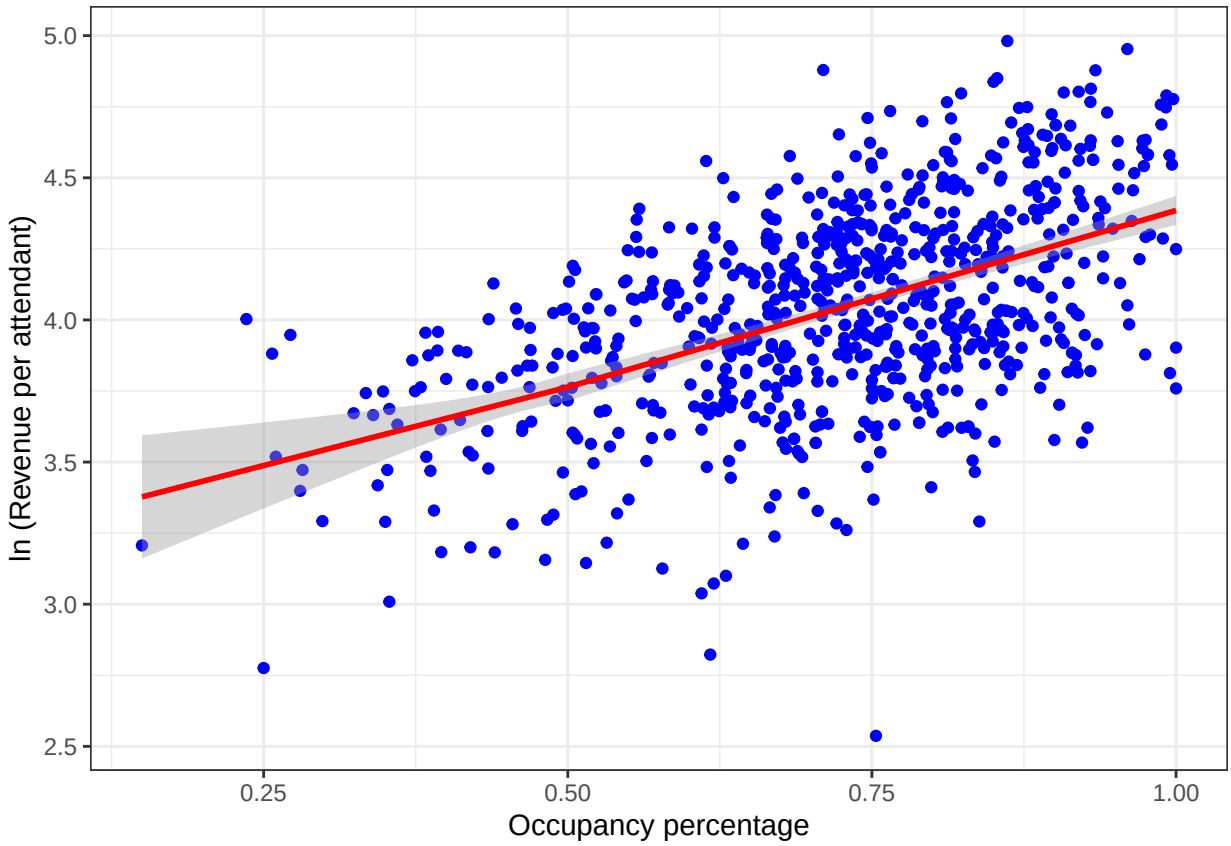
```
##
## Call:
## lm_robust(formula = ln_revenue_per_att ~ capacity_filled + capacity_filled_sq,
##     data = df)
##
## Standard error type:  HC2
##
## Coefficients:
##                    Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper  DF
## (Intercept)          3.3176     0.1654  20.064 1.327e-72   2.9931    3.642 785
## capacity_filled      0.7177     0.5023   1.429 1.535e-01  -0.2684    1.704 785
## capacity_filled_sq   0.3765     0.3715   1.013 3.111e-01  -0.3527    1.106 785
##
## Multiple R-squared:  0.2664 ,    Adjusted R-squared:  0.2645
## F-statistic: 145.4 on 2 and 785 DF,  p-value: < 2.2e-16
```



Regressipn 3 - Piecewise linear spline regression

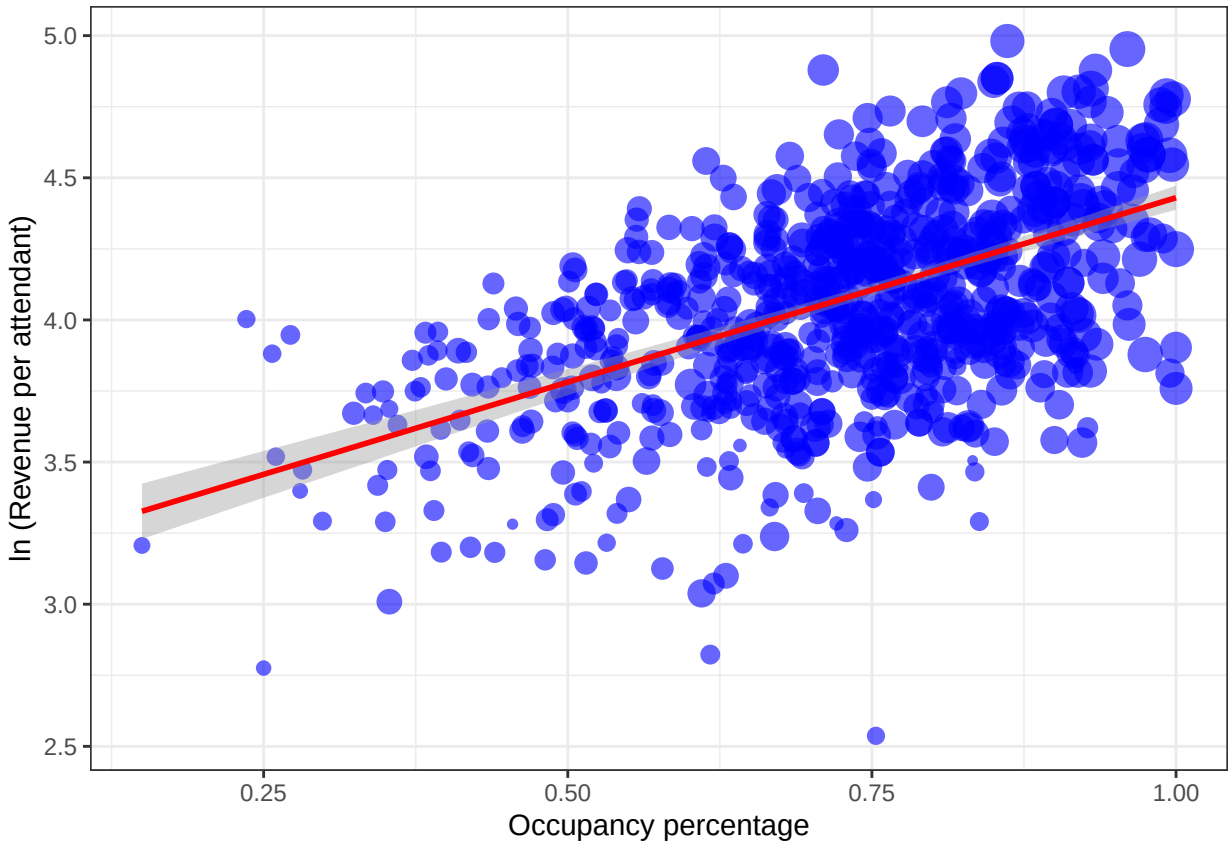
Using 0.5 as a cutof point

```
##
## Call:
## lm_robust(formula = ln_revenue_per_att ~ lspline(capacity_filled,
##     cutoff), data = df)
##
## Standard error type:  HC2
##
## Coefficients:
##                                  Estimate Std. Error t value  Pr(>|t|)
## (Intercept)                         3.212    0.17019   18.87 8.454e-66
## lspline(capacity_filled, cutoff)1   1.103    0.36171    3.05 2.362e-03
## lspline(capacity_filled, cutoff)2   1.243    0.09032   13.76 9.248e-39
##                                  CI Lower CI Upper  DF
## (Intercept)                        2.8779    3.546 785
## lspline(capacity_filled, cutoff)1  0.3934    1.813 785
## lspline(capacity_filled, cutoff)2  1.0657    1.420 785
##
## Multiple R-squared:  0.2655 ,    Adjusted R-squared:  0.2636
## F-statistic: 145.3 on 2 and 785 DF,  p-value: < 2.2e-16
```



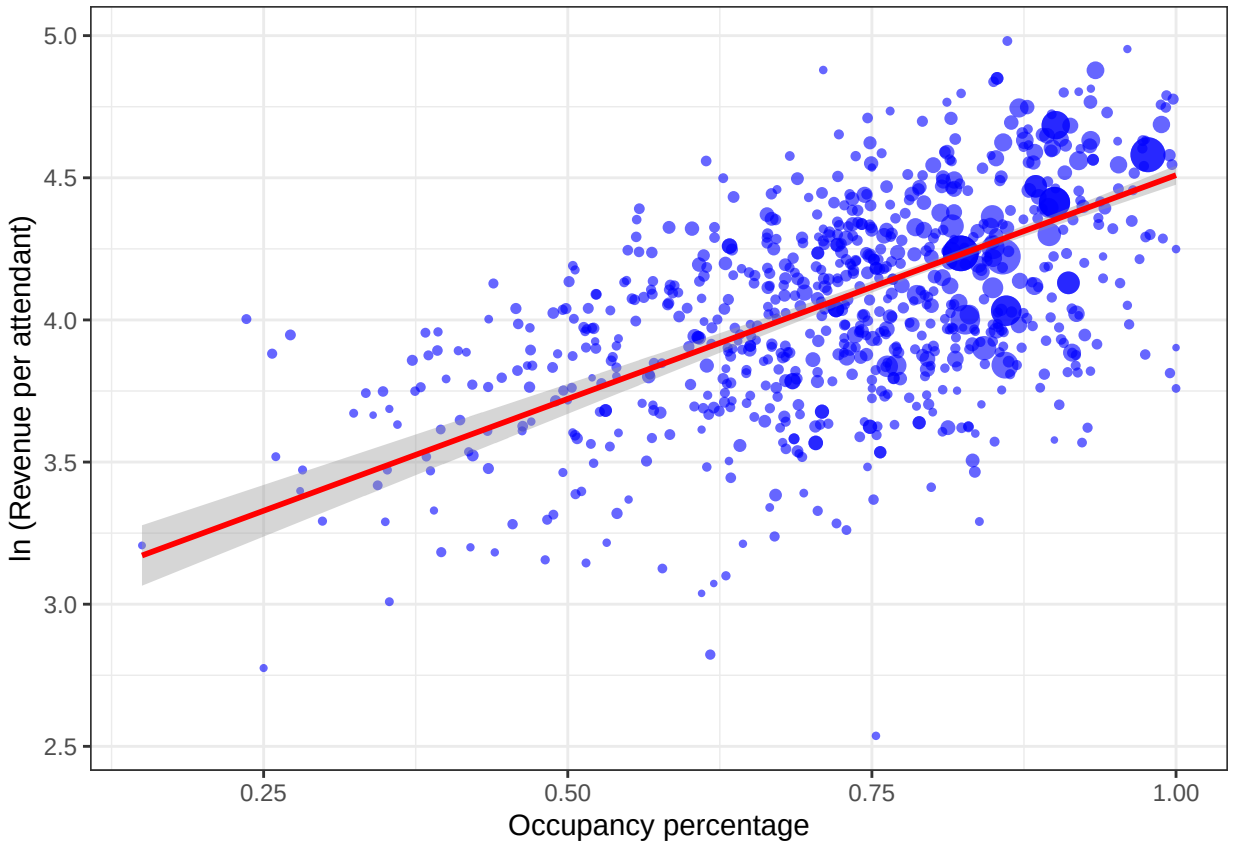
Regression 4 - Weighted linear regression, where weights = percentage of total revenue

```
##
## Call:
## lm_robust(formula = ln_revenue_per_att ~ capacity_filled, data = df,
##     weights = percentage_of_poss_profit)
##
## Weighted, Standard error type:  HC2
##
## Coefficients:
##                 Estimate Std. Error t value   Pr(>|t|) CI Lower CI Upper  DF
## (Intercept)        3.132    0.05713   54.83 8.233e-271    3.020    3.244 786
## capacity_filled    1.298    0.07849   16.53  6.364e-53    1.144    1.452 786
##
## Multiple R-squared:  0.2571 ,    Adjusted R-squared:  0.2561
## F-statistic: 273.4 on 1 and 786 DF,  p-value: < 2.2e-16

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



Regression 5 - Weighted linear regression, where weights = number of performances

```
##
## Call:
## lm_robust(formula = ln_revenue_per_att ~ capacity_filled, data = df,
##     weights = num_of_performances)
##
## Weighted, Standard error type:  HC2
##
## Coefficients:
##                 Estimate Std. Error t value   Pr(>|t|) CI Lower CI Upper  DF
## (Intercept)        2.935    0.09171   32.01 1.485e-144    2.755    3.115 786
## capacity_filled    1.574    0.12362   12.73  6.777e-34    1.331    1.817 786
##
## Multiple R-squared:  0.3303 ,    Adjusted R-squared:  0.3294
## F-statistic: 162.1 on 1 and 786 DF,  p-value: < 2.2e-16

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



Model Comparison

```
## The table was written to the file '/Users/Terez/OneDrive - Central European University/Data_Analysis_
```

Looks like these models with mainly one variable are not a great fit for the data. Therefore, I will include additional variables to try and get a better fit. Further, it looks like the original use of "Number of Performances" has no impact so I will try and create a dummy variable and use that instead. I will use 0 for any show that had less than one year of performances so less than 8*52 (416) and one for those that have had more.

Additional models

Check if it becomes better if one of the weights are included as variables

```
##
## Call:
## lm_robust(formula = ln_revenue_per_att ~ capacity_filled + percentage_of_poss_profit,
##     data = df, se_type = "HC2")
##
## Standard error type:  HC2
##
## Coefficients:
##                           Estimate Std. Error t value   Pr(>|t|) CI Lower
## (Intercept)                 3.3273    0.05130  64.865 1.380e-317  3.22661
```

```
## capacity_filled             0.2832    0.10105   2.802  5.203e-03  0.08479
## percentage_of_poss_profit   0.9884    0.08146  12.133  3.569e-31  0.82846
##                           CI Upper  DF
## (Intercept)                 3.4280 785
## capacity_filled             0.4815 785
## percentage_of_poss_profit   1.1483 785
##
## Multiple R-squared:  0.3874 ,    Adjusted R-squared:  0.3858
## F-statistic: 233.6 on 2 and 785 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm_robust(formula = ln_revenue_per_att ~ capacity_filled + as.factor(num_of_performances_d),
##     data = df, se_type = "HC2")
##
## Standard error type:  HC2
##
## Coefficients:
##                                  Estimate Std. Error t value   Pr(>|t|)
## (Intercept)                        3.1918    0.05286  60.387 2.877e-297
## capacity_filled                    1.1498    0.07526  15.279  2.484e-46
## as.factor(num_of_performances_d)1  0.1128    0.02769   4.072  5.140e-05
##                                  CI Lower CI Upper  DF
## (Intercept)                        3.0881   3.2956 785
## capacity_filled                    1.0021   1.2976 785
## as.factor(num_of_performances_d)1  0.0584   0.1671 785
##
## Multiple R-squared:  0.2778 ,    Adjusted R-squared:  0.276
## F-statistic: 170.2 on 2 and 785 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm_robust(formula = ln_revenue_per_att ~ capacity_filled + percentage_of_poss_profit +
##     as.factor(num_of_performances_d), data = df, se_type = "HC2")
##
## Standard error type:  HC2
##
## Coefficients:
##                                  Estimate Std. Error t value   Pr(>|t|)
## (Intercept)                        3.3460    0.05190  64.470 1.289e-315
## capacity_filled                    0.2567    0.10065   2.551  1.094e-02
## percentage_of_poss_profit          0.9652    0.08123  11.881  4.732e-30
## as.factor(num_of_performances_d)1  0.0743    0.02500   2.971  3.056e-03
##                                  CI Lower CI Upper  DF
## (Intercept)                       3.24410   3.4479 784
## capacity_filled                   0.05914   0.4543 784
## percentage_of_poss_profit         0.80569   1.1246 784
## as.factor(num_of_performances_d)1 0.02521   0.1234 784
##
## Multiple R-squared:  0.3927 ,    Adjusted R-squared:  0.3904
## F-statistic: 172.2 on 3 and 784 DF,  p-value: < 2.2e-16
```

```
##
```

```
## Call:
## lm_robust(formula = ln_revenue_per_att ~ capacity_filled + percentage_of_poss_profit +
##     as.factor(num_of_performances_d) + as.factor(show_type),
##     data = df, se_type = "HC2")
##
## Standard error type:  HC2
##
## Coefficients:
##                                   Estimate Std. Error t value   Pr(>|t|)
## (Intercept)                        3.42570    0.05297 64.6667 4.715e-316
## capacity_filled                    0.23419    0.09791  2.3919  1.700e-02
## percentage_of_poss_profit          0.97896    0.08005 12.2290  1.361e-31
## as.factor(num_of_performances_d)1  0.02045    0.02689  0.7606  4.471e-01
## as.factor(show_type)Play          -0.10331    0.02173 -4.7549  2.363e-06
## as.factor(show_type)Special       -0.08564    0.06654 -1.2870  1.985e-01
##                                   CI Lower CI Upper  DF
## (Intercept)                        3.32171  3.52969 782
## capacity_filled                    0.04199  0.42638 782
## percentage_of_poss_profit          0.82182  1.13610 782
## as.factor(num_of_performances_d)1 -0.03234  0.07324 782
## as.factor(show_type)Play          -0.14596 -0.06066 782
## as.factor(show_type)Special       -0.21627  0.04498 782
##
## Multiple R-squared:  0.4085 ,    Adjusted R-squared:  0.4048
## F-statistic: 112.1 on 5 and 782 DF,  p-value: < 2.2e-16
```

Explore again

```
## The table was written to the file '/Users/Terez/OneDrive - Central European University/Data_Analysis
```