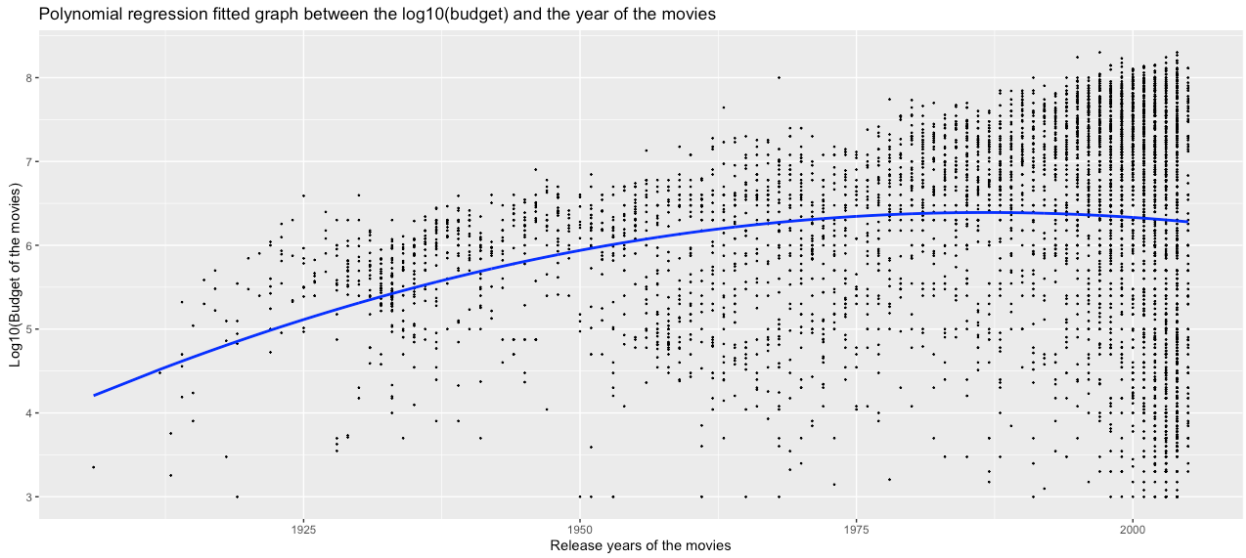
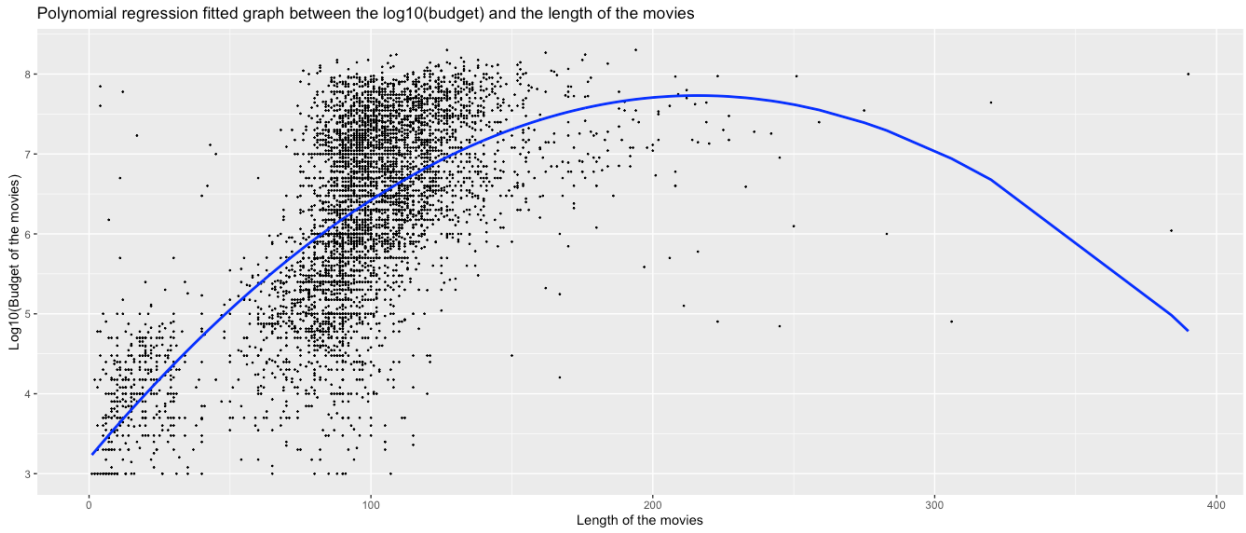
RITWIK BUDHIRAJA 2000916462 rbudhira@iu.edu

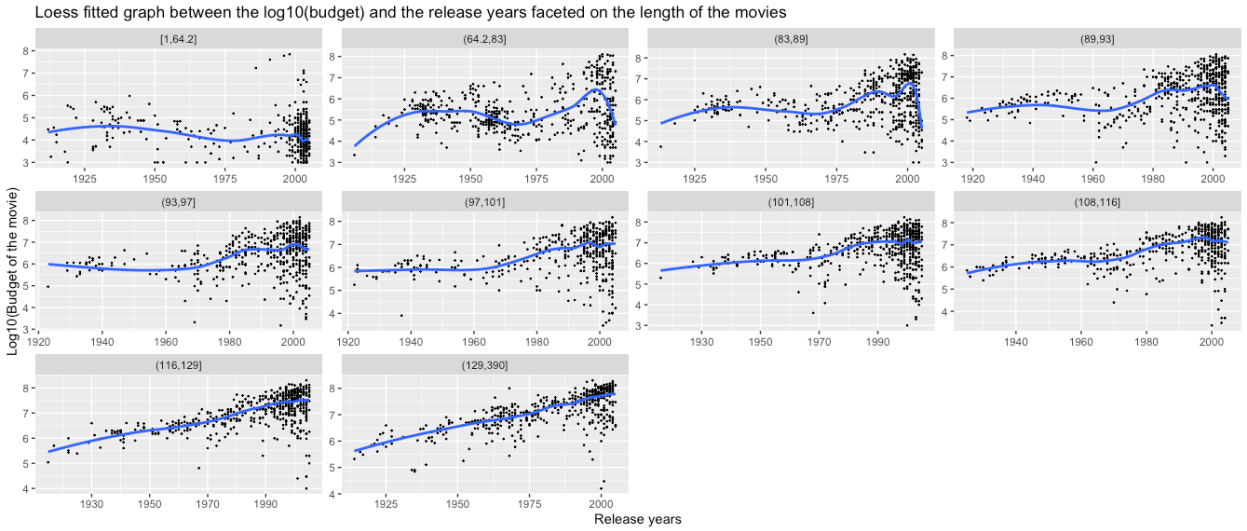
EDA ASSIGNMENT – 4

APPENDIX: I am attaching some other graphs for reference purposes.





RITWIK BUDHIRAJA 2000916462 rbudhira@iu.edu



ABSTRACT - This report talks about performing Exploratory Data Analysis on the open-sourced 'movie_budgets' data using R. The aim is to find trends and meaningful insights by creating data visualizations, transformations, and regression models amongst other methods.

KEYWORDS— movie_budgets, length, year, log10(budget), regressions, visualizations, correlations, effects, R.

QUESTION 1:

INTRODUCTION: Since the data is not particularly linear, I have used a loess model with degree 2 to depict the relationship between the log10(budget) and the release year of the movies, and between the log10(budget) and the length of the movies. Before finalizing on a loess model with degree 2, I tried to build a linear model, but since the relationship wasn't linear, it wasn't a good fit. Furthermore, I also tried to visualize a polynomial fit of degree 2, but that wasn't mapping well to the data.

LINE OF CODE TO PRODUCE THE MODEL: loess_length = loess(log10(budget) ~ length, span = 0.5, degree = 2, data = data, family = "symmetric")

- Should you fit a linear or curved function for year?

  I have fit a curved function for year because the relationship between the log10(budget) and the release year of the movies was somewhat linear, but there were some sub-data concentrations throughout the graph, so a loess model of degree 2 was mapping that pretty well.

- Should you fit a linear or curved function for length?

  I have fit a curved function for length because the relationship between the log10(budget) and the length of the movies wasn't linear.

- Do you need an interaction between year and length?

  On using the ggpairs function, I found that the correlation between year and length was relatively low, so I decided to not include an interaction between the two.

- What span should you use in your loess smoother?

  I have used span = 0.4 for year because since there were clusters in the plot, taking span = 0.75/1 was being more generalized and hence smoother. For length, I have taken span = 0.5 for the same reason, that is to avoid generalizing and smoothening.

- Should you fit using least squares or a robust fit?

  I have used family = "symmetric" to use robust fit. This was done in order to decrease the impact of outliers on the model.

QUESTION 2:

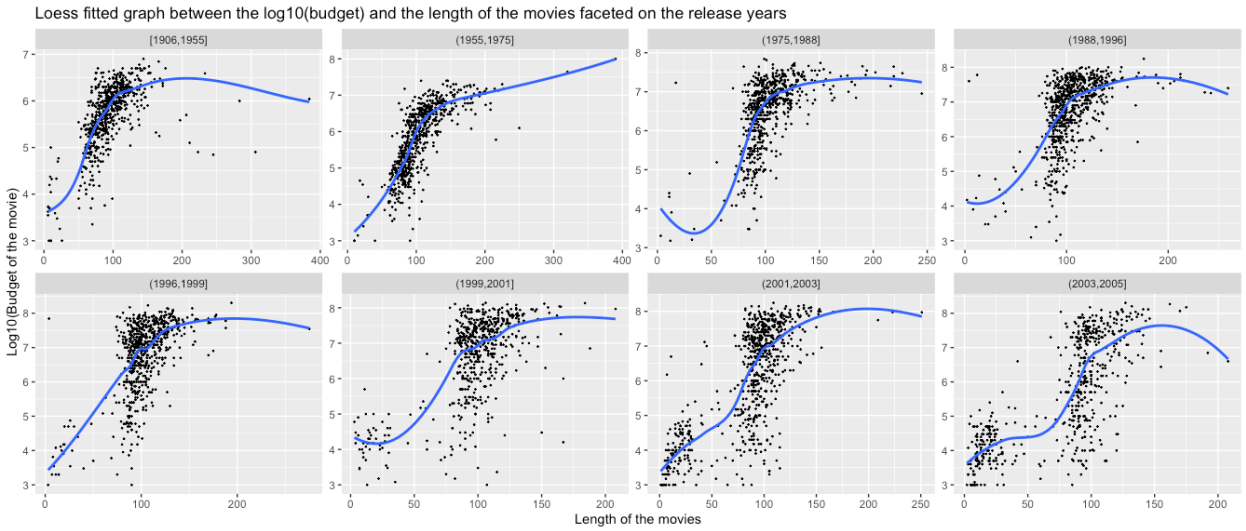
INTRODUCTION: I have drawn a set of faceted plots to display my loess fit between the log10(budget) and the length of the movies, conditioned on the release years. For this plot, I divided the year column into 8 intervals.

LINE OF CODE:

```
# Loess fitted graph between the log10(budget) and the length of the movies faceted on the release years
ggplot(data, aes(x = length, y = log10(budget))) +
  geom_point(size = 0.2) +
  facet_wrap(~ cut_number(year, n = 8), ncol = 4, scales = "free") +
  ylab("Log10(Budget of the movie)") +
  xlab("Length of the movies") +
  geom_smooth(method = "loess", se = FALSE, span = 0.5, method.args = list(degree = 2, family = "symmetric")) +
  ggtitle("Loess fitted graph between the log10(budget) and the length of the movies faceted on the release years")
```

RITWIK BUDHIRAJA 2000916462 rbudhira@iu.edu

PLOT:



- The relationship between the log10(budget) and the length of the movies is somewhat linear minus the impact of the outliers. I have, personally, used a quadratic model to also consider the outliers (but only to a certain extent).
- The variance is similar across all the graphs.
- The faceted graphs also seem to be a good approximation of the unfaceted graph.

QUESTION 3:

INTRODUCTION: In this graph, I've used a raster-and-contour plot to depict the impact of both length and year on the budget of the movies. For this plot, I have used 15 bins to divide the fitted budget values. I cannot deduce any extra information from this plot other than the plot in question 2.

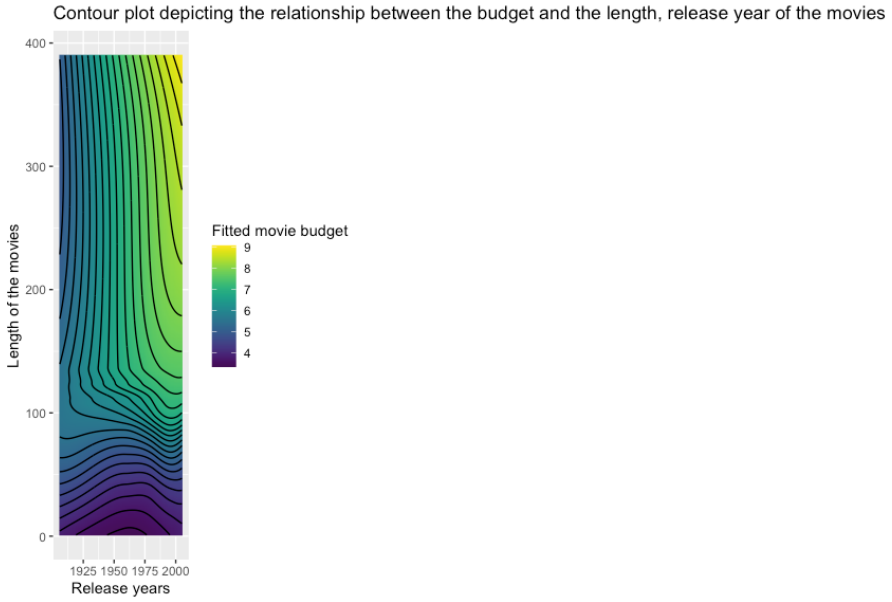
LINE OF CODE:

```
# Creating x and y dimensions
movie_grid = data.frame(expand.grid(
  length = seq(1, 390, 1),
  year = seq(1906, 2005, 1)))

# Building a loess model between the budget and the length, release year of the movies
loess_ques3 = loess(log10(budget) ~ year * length,
                    data = data, span = 0.5, degree = 2,
                    family = "symmetric", normalize = FALSE)
```

RITWIK BUDHIRAJA 2000916462 rbudhira@iu.edu

```
loess_contour = augment(loess_ques3, newdata = movie_grid)

ggplot(loess_contour, aes(x = year, y = length, fill = .fitted, z = .fitted)) +
  geom_raster() +
  geom_contour(bins = 30, color = "black") +
  coord_fixed() +
  scale_fill_viridis("Fitted movie budget") +
  ylab("Length of the movies") +
  xlab("Release years") +
  ggtitle("Contour plot depicting the relationship between the budget and the length, release year of the movies")
```

PLOT:



- For the range of length 200-400, the budget of the movie is increasing rapidly up till about 1980. This can be commented upon because of the fact that the contours in that range are closely placed, and so the slope in that region must be really high. This could be accredited to inflation or the modernizing industry. After that, the rate slows down as the gaps between the contours widen.

- A similar effect can be seen in the movies of length 50-100. Around the year 2000, the budget of the movie could be rapidly increasing and then slows down after crossing the length 100. This could be because of the increasing demand for shorter length movies during those years.