Tuition prediction modeling

Liza Luizova & Maya McNeill

December 7, 2025

Student ID's:

- Liza Luizova (169089763)
- Maya McNeill (169077416)

```
library(tidyverse)
library(readr)
library(patchwork)
library(ggridges)
library(janitor)
theme_set(theme_bw())
knitr::opts_chunk$set(
  echo = TRUE,     # show code
  message = FALSE, # hide package startup messages
  warning = FALSE  # hide warnings
) # Prevents code from showing in the pdf.

education_costs <- read_csv(
  "C:/Users/  /Desktop/DATA100 R/DATA100-FINAL-PROJECT/International_Education_Costs.csv",
  show_col_types = FALSE)
```

Introduction

This project uses the International Education Costs dataset, which contains information about the financial aspects of studying abroad. The dataset includes variables such as country, city, university, program level, tuition fees (in USD), rent, insurance, visa fees, exchange rates, living cost index, and program duration. These variables provide a detailed view of both academic and living expenses faced by international students and allow for comparisons of education costs across countries and institutions.

Goals

The goal of this analysis is to explore and model the factors that influence tuition costs for international students. In particular, this project focuses on understanding how tuition varies with living cost index, program duration, and geographic region.

Basic Data Cleaning

The current dataset seems to be usable, but not fully ready for modeling. Several cleaning steps are necessary to ensure that the data is tidy, consistent and suitable for analysis.

First of all, we need to make sure that our variables have appropriate data types: quantitative should be numeric, qualitative should be factor. We selected these variables and mutated them into an appropriate data type.

Second of all, Character columns sometimes contained unnecessary white space (leading or trailing spaces). All character variables were cleaned using str_trim() to ensure consistency when grouping or filtering.

Third of all, the dataset was inspected for missing values. Rows with missing values in essential cost related variables (tuition, rent, insurance, visa fees) were removed, as they would prevent accurate modeling.

Lastly, we created a couple of new variables to better support modeling and analysis: Annual rent (annual_rent), Total annual cost (total_annual_cost) and Local currency tuition with rent (was adjusted based on exchange currency) - these variables can improve predictive performance and provide more meaningful insights.

For further analysis and visualization, we also reshaped cost related variables using pivot_longer(), resulting in cost_type and cost_value columns. It will be useful for comparing cost categories across countries or institutions.

```
# Cleaning before modeling
education_cleaned <- education_costs |>
  clean_names() |>
  mutate(across(where(is.character), str_trim)) |>
  mutate(
    country = factor(country),
    city  = factor(city),
    university = factor(university),
    program = factor(program),
    level = factor(level)
  ) |>
  mutate(
    tuition_usd = as.numeric(tuition_usd),
    rent_usd  = as.numeric(rent_usd),
    insurance_usd = as.numeric(insurance_usd),
    visa_fee_usd = as.numeric(visa_fee_usd),
    exchange_rate = as.numeric(exchange_rate)
  ) |>
  mutate(
    annual_rent = rent_usd * 12,
    total_annual_cost = tuition_usd + annual_rent + insurance_usd + visa_fee_usd,
    tuition_local = tuition_usd / exchange_rate,
    rent_local  = rent_usd / exchange_rate
  ) |>
  filter(
    !is.na(tuition_usd),
    !is.na(rent_usd),
    !is.na(insurance_usd),
    !is.na(visa_fee_usd)
  )

education_long <- education_cleaned |>
```

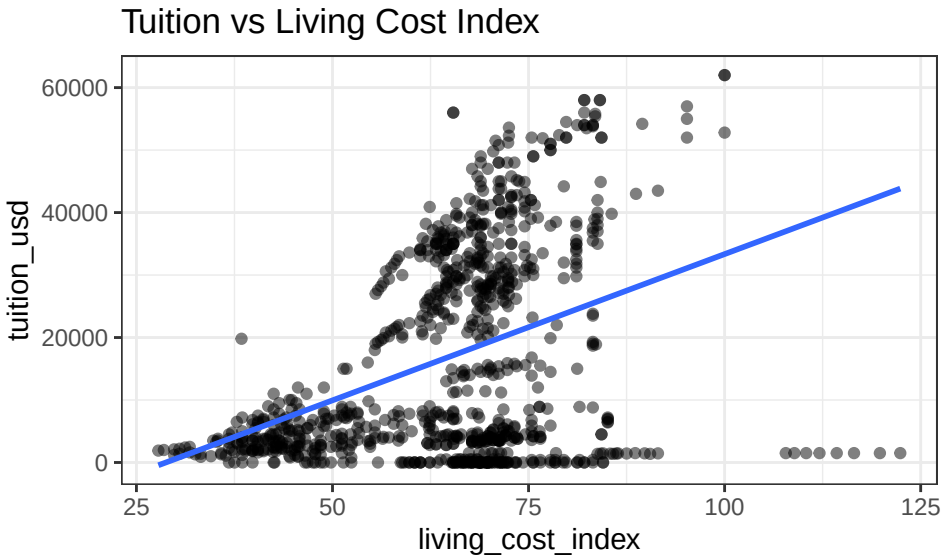
```
  pivot_longer(
    cols = c(tuition_usd, rent_usd, insurance_usd, visa_fee_usd),
    names_to = "cost_type",
    values_to = "cost_value"
  )
```

Exploratory Plots

Visualization is an essential part of the Exploratory Data Analysis. Here, we will explore how the tuition varies and relates to other factors that could impact it, by using 3 plots.

Plot 1: How does tuition relate to the living cost index?

```
plot1 <- education_cleaned |>
  ggplot(aes(x = living_cost_index, y = tuition_usd)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Tuition vs Living Cost Index")
plot1
```



- Countries with higher living costs generally appear to have higher tuition fees. This suggests that cost of living and tuition may be influenced by similar economic factors.
- Although the trend is positive, the points are widely scattered around the trend line. Living cost index alone does not strongly predict tuition. There are likely other important variables influencing tuition
- High tuition but moderate living costs, and Low tuition in countries with relatively high living costs.

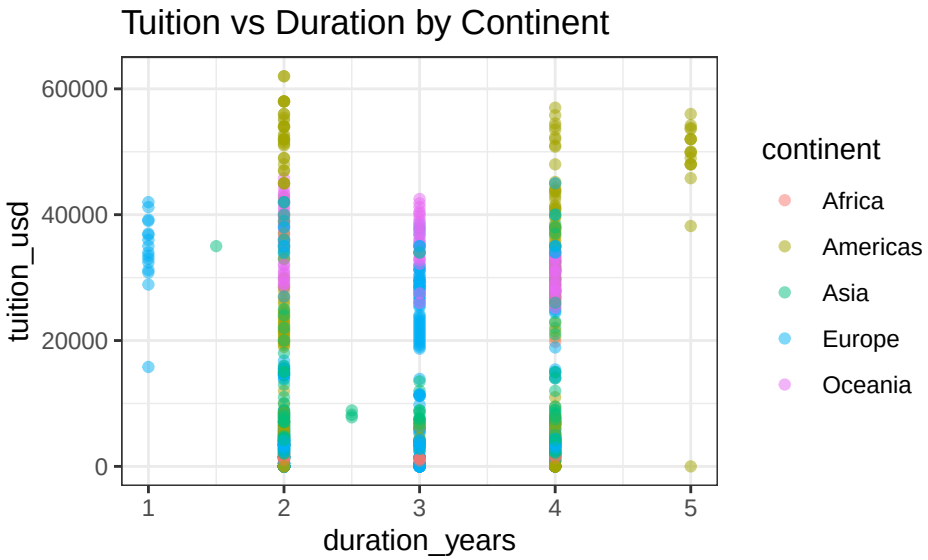
Here, we can conclude that this plot shows a large variation. The points of data look grouped in particular places, and the distance over the "best-fit" line does not really seem to be symmetric. Thus, adding other factors/predictors could improve explaining the differences in tuition.

Plot 2: Distribution of Tuition across programs by countries

```
library(countrycode)

education_cleaned <- education_cleaned |>
  mutate(
    continent = countrycode(country, "country.name", "continent")
  )

p3 <- education_cleaned |>
  ggplot(aes(x = duration_years, y = tuition_usd, color = continent)) +
  geom_point(alpha = 0.5) +
  labs(title = "Tuition vs Duration by Continent")
p3
```



- There is significant overlap between 1-year and 2-year programs.
- Continent Europe: Mostly lower-tuition programs, even longer programs remain relatively affordable, points cluster in the lower tuition range compared to other continents.
- Continent Asia: A broad range of durations, tuition varies widely but tends to sit in the mid-range, some outliers with higher tuition.
- Continents Americas: Typically higher tuition, programs with similar duration cost much more than in Europe or Asia, smaller spread in duration (mostly 1–2 year programs).
- Continent Oceania: very similar to Americas, but with moderate number of long duration programs.
- Continent Africa: generally have lower tuition, and there is a limited variation between duration and tuition.

Based on this plot, we can conclude that has a weaker linear relationship with tuition, as tuition does not consistently increase with duration_years and there are some overlaps between different years of studying.

Exploratory Modeling: Testing

As we move forward with creating a suitable model for predicting the tuition, the dataset was divided into separate subsets to allow for proper training and evaluation. Splitting the data helps ensure that the model is assessed on unseen data and reduces the risk of overfitting.

```
set.seed(123)
n <- nrow(education_cleaned)
train_idx <- sample(seq_len(n), size = 0.6 * n)
remaining_idx <- setdiff(seq_len(n), train_idx)
valid_idx <- sample(remaining_idx, size = 0.5 * length(remaining_idx))
test_idx  <- setdiff(remaining_idx, valid_idx)

# data sets
train_data <- education_cleaned[train_idx, ]
valid_data <- education_cleaned[valid_idx, ]
test_data  <- education_cleaned[test_idx, ]
```

Here, we broke down our data set into a standard 60% training, 20% validating and 20% evaluating.

Exploratory Modeling: Polynomial regression modeling and choosing the right degree

Polynomial regression was used to model potential nonlinear relationships between living_cost_index and tuition_usd. Different polynomial degrees were tested to find the model that balanced flexibility and generalization performance.

The model was evaluated using Root Mean Squared Error (RMSE) on the validation dataset. The degree that produced the lowest validation RMSE was selected as the optimal polynomial complexity.

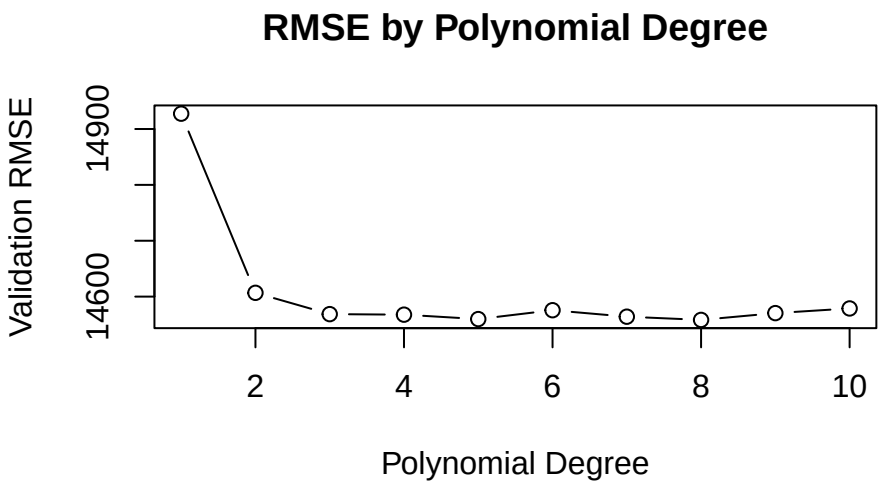
```
set.seed(123)

degrees <- 1:10

rmse_values <- numeric(length(degrees))

for (d in degrees) {
  model <- lm(tuition_usd ~ poly(living_cost_index, d, raw = TRUE),
              data = train_data)
  preds <- predict(model, newdata = valid_data)
  rmse_values[d] <- sqrt(mean((valid_data$tuition_usd - preds)^2))
}
best_degree <- degrees[which.min(rmse_values)]

plot(degrees, rmse_values, type = "b",
     xlab = "Polynomial Degree",
     ylab = "Validation RMSE",
     main = "RMSE by Polynomial Degree")
```



- The plot seems to show the absolute minimum values 5 and 8 from this range. Degree 5 is preferred because it achieves nearly the same RMSE as degree 8 but with a simpler curve and lower overfitting risk.

Model testing and visualization check

After selecting the polynomial degree, the training and validation datasets were combined to build a final model using the full development dataset. This model was then evaluated on the held-out test set to obtain unbiased estimates of predictive performance.

Key performance metric that were calculated: RMSE (Root Mean Squared Error) to measure overall prediction error, MAE (Mean Absolute Error) to assess average absolute prediction deviation, "R squared" (coefficient of determination) to quantify how much variance in tuition is explained by the model.

```
train_all <- bind_rows(train_data, valid_data)

final_model <- lm(tuition_usd ~ poly(living_cost_index, 5, raw = TRUE),
                  data = train_all)

pred_test <- predict(final_model, newdata = test_data)
rmse_test <- sqrt(mean((test_data$tuition_usd - pred_test)^2))
mae_test  <- mean(abs(test_data$tuition_usd - pred_test))
r2_test   <- 1 - sum((test_data$tuition_usd - pred_test)^2) /
  sum((test_data$tuition_usd - mean(test_data$tuition_usd))^2)

cat("Test RMSE:", rmse_test, "\nTest MAE:", mae_test, "\nTest R2:", r2_test, "\n")
```

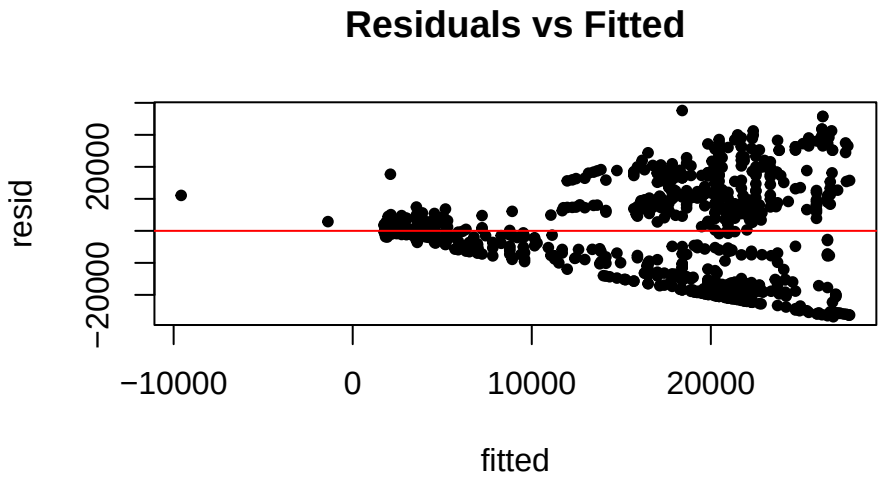
```
## Test RMSE: 14597.55
## Test MAE: 12019.83
## Test R2: 0.1779541
```

```r
pred_train <- predict(final_model, newdata = train_all)
rmse_train <- sqrt(mean((train_all$tuition_usd - pred_train)^2))
cat("Train RMSE:", rmse_train, " | Test RMSE:", rmse_test, "\n")
```

```
## Train RMSE: 14864.18  | Test RMSE: 14597.55
```

```r
resid <- resid(final_model)
fitted <- fitted(final_model)

plot(fitted, resid, pch=20, main="Residuals vs Fitted")
abline(h=0, col="red")
```




```r
shapiro.test(sample(resid, min(length(resid), 5000)))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  sample(resid, min(length(resid), 5000))
## W = 0.97336, p-value = 3.228e-10
```

We can observe that our Test RMSE is not by much less than Train RMSE => no signs of overfitting. Even though there might be other factors that could lead to getting the random or systematic error between those 2 values, we can conclude that this model generalizes pretty well.

Summary table with performance variables

To improve interpretation and presentation in the final report, the model performance metrics were summarized in a formatted table. This table provides a compact overview of prediction accuracy and allows easy comparison of RMSE, MAE, and R squared values.

```r
library(kableExtra)

model2 <- lm(tuition_usd ~ poly(living_cost_index, 5, raw=TRUE) +
               duration_years + continent, data=train_data)

preds <- predict(model2, newdata = test_data)
rmse <- sqrt(mean((test_data$tuition_usd - preds)^2))
mae  <- mean(abs(test_data$tuition_usd - preds))
r2   <- 1 - sum((test_data$tuition_usd - preds)^2) /
            sum((test_data$tuition_usd - mean(test_data$tuition_usd))^2)

model_results <- tibble(
  Metric = c("RMSE", "MAE", "R-squared"),
  Value  = c(rmse, mae, r2)
)
model_results |>
  knitr::kable(
    caption = "Model Performance Metrics",
    digits = 3,
    align = "c"
  ) |>
  kable_styling(
    full_width = FALSE,
    position = "center"
  ) |>
  row_spec(0, bold = TRUE) |>
  column_spec(1:2, border_left = TRUE, border_right = TRUE)
```

Table 1: Model Performance Metrics

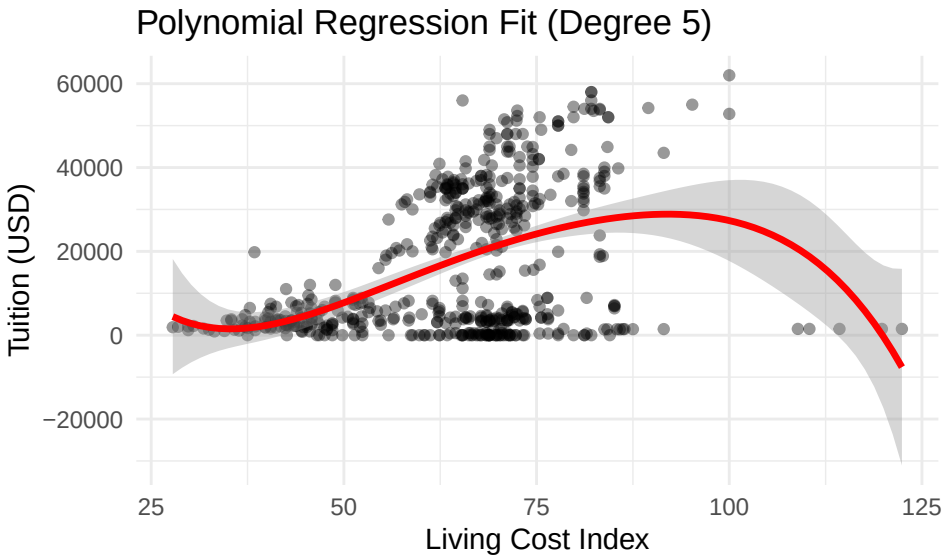
Metric	Value
RMSE	10914.562
MAE	9026.892
R-squared	0.540

Polynomial regression curve

A final visualization was created by plotting the observed tuition values against the living cost index, overlaid with the fitted degree 5 polynomial regression curve. This figure visually demonstrates the nonlinear relationship captured by the model and shows how the fitted curve follows the general trend of the data while smoothing random noise.

```r
ggplot(train_data, aes(x = living_cost_index, y = tuition_usd)) +
  geom_point(alpha = 0.4) +
  stat_smooth(
    method = "lm",
    formula = y ~ poly(x, 5, raw = TRUE),
    color = "red",
    linewidth = 1.2
  ) +
```

```
  labs(
    title = "Polynomial Regression Fit (Degree 5)",
    x = "Living Cost Index",
    y = "Tuition (USD)"
  ) +
  theme_minimal()
```



- The red curve shows a non-linear relationship between living cost index and tuition.
- Tuition generally increases as living costs rise, but not in a strictly linear way.
- The curve captures subtle bends in the trend, suggesting that a polynomial model fits the structure of the data better than a straight line.

Overall, the plot suggests that the 5th-degree polynomial regression model captures the main trend in tuition reasonably well. The data points appear to be fairly symmetrically distributed around the fitted curve, although greater dispersion is observed at higher values of the living cost index, indicating increased variability in tuition for more expensive countries.

Conclusion

This project explored the financial structure of international education by cleaning, visualizing, and modeling a dataset of education-related costs across multiple countries and institutions. After extensive preparation, several exploratory analyses revealed meaningful relationships between tuition fees, living cost indices, program duration, and geographic region. Polynomial regression models were tested and optimized to capture non-linear trends in the data, and model performance was evaluated using standard validation techniques. The results showed that living costs, duration of study, and continent play important roles in explaining variation in tuition fees, while also highlighting the limits of prediction due to structural differences across institutions and countries.

Overall, this study demonstrates the value of combining careful data cleaning, exploratory visualization, and advanced modeling techniques to better understand the economic factors that influence international education costs. Future work could expand this analysis by incorporating additional institutional features such as university rankings or funding structures to further improve predictive accuracy.

References

1) Dataset:

Shamim, A. (2025, May 7). Cost of international education. Kaggle.
https://www.kaggle.com/datasets/adilshamim8/cost-of-international-education/data

2) Library countrycode: sort the countries by continents:

DataCamp. (n.d.). Countrycode: Convert Country Codes. RDocumentation.
https://www.rdocumentation.org/packages/countrycode/versions/1.6.1/topics/countrycode

3) Splitting the data:

DataCamp. (n.d.-b). Splitting the data set. Splitting the data set | R.
https://campus.datacamp.com/courses/credit-risk-modeling-in-r/chapter-1-introduction-and-data-preprocessing?ex=12

4) Normality test for performance variables:

Statistical tools for high-throughput data analysis. (n.d.). Normality test in R. STHDA.
https://www.sthda.com/english/wiki/normality-test-in-r

5) How to determine the optimal degree for polynomial regression:

Clark Science Center. (n.d.). 7.8.1 polynomial regression and step functions¶. Lab 12 - Polynomial Regression and Step Functions in R. https://www.science.smith.edu/~jcrouser/SDS293/labs/lab12-r.html

6) KableExtra: formatting tables when knitting:

Zhu, H. (2024, January 23). Create awesome HTML table with knitr::kable and kableextra.
https://cran.r-project.org/web/packages/kableExtra/vignettes/awesome_table_in_html.html

7) The meaning of Training RMSE value being less than the Test one:

Malato, G. (2021, November 8). Why training set should always be smaller than test set. Your Data Teacher.
https://www.yourdatateacher.com/2021/05/03/why-training-set-should-always-be-smaller-than-test-set/

8) Plotting polynomial regression with ggplot:

Neitmann, T. (2021, October 3). How to add a regression line to a ggplot?. Thomas' adventuRe.
https://thomasadventure.blog/posts/ggplot-regression-line/