Understanding the Future of Higher Ed Landscape in Dubai

Klim Popov

23 May, 2020

1 Abstract

The market for higher education institutions in the UAE is fearce. Due to to the recent pandemic of COVID-19, many institutions are expected to partly open their educational programs online, which expected to result in overcrowding the market with educational programs supply.

In this project we will discuss several tools to support university decisions on actions required to be taken amid the aftermath of the pandemic.

This project aims to discover the current landscape of Higher Ed in Dubai and analyze several forecast models to support decision making processes in the universities.

The project includes analysis on three groups of collected data: from government entity (KHDA census), Google Reviews and Google Trends. The first dataset analysis will include Linear Model, Quantile Regression Model, Random Forest Model. The second dataset will include Factorization Model and Sentiment Analysis. The last dataset will include time-series analysis and trend prediction with Holt-Winters.

The datasets were divided into train and test sets as per the requirement for this project.

This report consists of the following parts: The overview section describes the dataset and summarizes the goal of the project and key steps that were performed; Methods section explains the process and techniques used, including data cleaning, data exploration and visualization, insights gained, and modeling approach for each of the performed analysis, followed by the Results section, subsequent to each analysis, that presents the modeling results and discusses the model performance for the Linear Model and Factorization Model; Finally, the Conclusion section provides a summary of the report, the limitations and future research.

Key findings: There are 10 key findings on the future of higher education in Dubai, outlined in the colclusion

Files attached to this report:

- RMD script
- R script

Contents

2 Overview of the project

2.1 Project Requirements

Submission Files:

The project requires PDF report, report in Rmd format and an R script

Report Structure and Content: The report documents the analysis and presents the findings, along with supporting statistics and figures. The following thorough explanation or justification for various steps of your project needs to be provided: why a specific train/test split (e.g. 50/50 vs 90/10) or algorithm was used. The report must be written in English and include at least the following sections:

- An introduction/overview/executive summary section that describes the dataset and variables, and summarizes the goal of the project and key steps that were performed.
- A methods/analysis section that explains the process and techniques used, including data cleaning, data exploration and visualization, any insights gained, and your modeling approach. At least two different models or algorithms must be used, with at least one being more advanced than simple linear regression for prediction problems.
- A results section that presents the modeling results and discusses the model performance.
- A conclusion section that gives a brief summary of the report, its potential impact, its limitations, and future work.

Code Requirements: The code in the R script should run without errors and should be well-commented and easy to follow. It should also use relative file paths and automatically install missing packages. The dataset you use should either be automatically downloaded by the code or provided in the GitHub repo:

- Code runs easily, is easy to follow, is consistent with the report, and is well-commented. All file paths are relative and missing packages are automatically installed with if(!require) statements.

2.2 Overview of used resources

The specifications of the R Studio session and libraries used in the report can be found here. In this project, we utilized R Server run on Google Cloud Platform. All file paths associated with the project are relative and missing packages are automatically installed with if(!require) statements. The packages are loaded through Pacman library in order to install only the packages which are not present on the machine, which runs the R/RMD script. It is possible to run both scripts to review the datasets and results. The functionality was tested with 2 independent machines.

2.3 Data Collection and Overview

The data collection sections describes 3 groups of sources for this project. The collected data was uploaded on github and corresponds to the following sections of our report:

- KHDA data: files HigherEdData.csv, Unis-names.csv - census of universities located in Dubai and licensed by the KHDA;
- Google Maps data: files googleUnis.csv and reviews.csv - information about Google profiles of the universities in Dubai and their reviews, was gathered through independent scraper for Google Reviews outScrapper, as it was not possible to gather the information through R directly due to limitations of JAVA installed on the machine;
- Google Trends data: historical data in form of timeseries for crawled keywords, via gtrendsR package.

2.3.0.1 University Census Set

The data was initially collected from an official source - Knowledge and Human Development Authority of Dubai with their annual census of institutions in Dubai. We have added information on rankings from the same portal which was not present in the original file.

While we specified the range from 2013 to 2019, for several institutions the census was not completed or completed partly. This provides a very big limitation for our analysis.

Ranking, curriculum, years in Dubai information was gathered from KHDA website, while ranking was adjusted to represent most popular (1-UK, US, Australia), Arab and European curriuclum (2-UAE, EU), and Asian curriculum (3 - India, Pakistan, non-GCC countries)

HigherEdData and Unis_names Sets

HigherEdData is structured as data frame with 1882 observation and 9 variables, **Unis_names** consists of 36 observations and 6 variables. The composition of the variables:

Table 1: Overview of variables in HigherEdData and Unis-names

Name	Comment	Dataset
Year	Year of census collected	HigherEdData
Uni	University Name	HigherEdData
Location	Location	HigherEdData
Lat	Lattitude of location	HigherEdData
Long	Longtitude of location	HigherEdData
Major	Major or specialization of students in census	HigherEdData
Level	Undergraduate or Postgraduate Students	HigherEdData
Status	Enrolled or Graduated	HigherEdData
Students	Number of students in the census	HigherEdData
UniId	University Id	Unis-names
University.Name	Name of the University	Unis-names
Avarage.Tuition.Fee	Avarege Tuition Fees	Unis-names
Ranking	Ranking of the university as of 2019	Unis-names
Curriculum.Rating	Rating of the curriculum depending on the country of origin	Unis-names
YearsinDubai	Number of years operating in Dubai	Unis-names

The data in both datasets appears to be not tidy enough to start our analysis. The below first six records from the dataset demonstrate that the data structure needs to be changed:

Especially, for the following observations in the first dataset:

- Not required data: Lat and Long
- Students presented as summative data. This will limit our model for analysis. There is no clear display of new, concurrent, withdrawals, transfers to make the sense of the data, therefore, only total enrolled numbers will be used
- No ID assigned to any of the entities

Table 2: First six records of HigherEdData

Year	Uni	Location	Lat	Long	Major	Level	Status	Students
2013	American University in Dubai	JLT	25.06942	55.14229	Architecture and Construction	Undergraduate	Enrolled	238
2013	American University in Dubai	JLT	25.06942	55.14229	Business	Undergraduate	Enrolled	1064
2013	American University in Dubai	JLT	25.06942	55.14229	Engineering	Undergraduate	Enrolled	190
2013	American University in Dubai	JLT	25.06942	55.14229	Foundation	Undergraduate	Enrolled	155
2013	American University in Dubai	JLT	25.06942	55.14229	Humanities	Undergraduate	Enrolled	78
2013	American University in Dubai	JLT	25.06942	55.14229	Information Technology	Undergraduate	Enrolled	4

- Dataset has information on Graduated students - not the focus of our study

And the second dataset:

Table 3: First six records of Unisnames

UniId	University.Name	Avarage.Tuition.Fee	Ranking	Curriculum.Rating	YearsinDubai
1	Abu Dhabi University	120000	5	1	25
2	American University in Dubai	120000	5	1	13
3	Amity University Dubai	40000	4	3	25
4	Birla Institute of Technology and Science Pilani (Bits Pilani) Dubai Campus	35000	3	3	15
5	British University in Dubai	60000	4	1	10
6	City University of London	80000	4	1	4

Especially, for the following observations in the second dataset:

- No ID assigned to any of the entities
- The name of the university most liklely to correspond to subsequent names in the first dataset.

Hence, prior to starting any analysis of the data, the dataset must be put in order.

2.3.0.2 University Review Data - Google Maps

The datasets **googleUni** and **review** provide information about Google profiles of the universities in Dubai and their reviews; the data was gathered through independent scraper for Google Reviews outScrapper. Prior to sharing the datasets, the respective names of google users who have rated the universities were deleted and anonymized.

GoogleUni and Review Sets

GUni is structured as data frame with 77 observation and 6 variables, **GReviews** consists of 3525 observations and 5 variables. The composition of the variables is presented in the overview table.

The data in both datasets appears to be not tidy enough to start our analysis. The below first six records from the dataset demonstrate that the data structure needs to be changed.

Especially, for the following observations in the first dataset:

Table 4: Overview of Guni and GReviews

Name	Comment	Dataset
name	Name of the University	GUni
latitude	Latitude of the location	GUni
longitude	Longtitude of the location	GUni
reviews	Total number of reviews	GUni
photos_count	Total number of photos assosiated with the reviews	GUni
google_id	Google Id of the university	GUni
google_id	Google Id of the university	GReviews
autor_id	Author Id of the review	GReviews
review_rating	Total score for rating (integer)	GReviews
review_timestamp	Timestamp in EPOCH format (seconds since Jan 1, 1970)	GReviews
review_text	Text of the review, translated text of the review	GReviews

Table 5: First six records of GUni

name	latitude	longitude	reviews	photos_count	google_id
Heriot-Watt University Dubai	25.12913	55.41649	137	87	0x3e5f638e7ac4f2d3:0xec919842ff503d0b
Middlesex University Dubai	25.10343	55.16434	96	87	0x3e5f6b6b695079cf:0x39837c77fd4e2851
Canadian University Dubai	25.20875	55.27034	178	90	0x3e5f426278f61349:0x5cc89b41e81e0384
University of Wollongong in Dubai (UOWD)	25.10370	55.16491	144	505	0x3e5f6b6b665006a3:0x84337ef2f5567aa5
University of Dubai	25.10609	55.41438	112	139	0x3e5f5cd08a0fa617:0xd2b8edf0881c4ecc
Amity University - Dubai Campus	25.12328	55.41939	163	493	0x3e5f639700000001:0x20597afbdab1b296

- Not required data: Lat and Long
- Google ID is very complex

And the second dataset:

Table 6: First six records of GReviews

google_id	autor_id	review_rating	review_timestamp	review_text
0x3e5f5d98991526f7:0x3b20c199e54f1db1	1.145546e+20	4	1545551548	One of the top-ranked universities in the world and in the region. Cass Business School is the ultimate place to receive an executive MBA. The weekend modular format of the program makes it easy to balance between work, life and study.
0x3e5f5d37a6e559bd:0xc72489cc2605d63	1.174627e+20	5	1564819159	Attended a CME by Cardiff University and Emirates Diabetes Society conducted at Mohammed Bin Rashid University of Medicine and health Sciences. The venue is excellent. The auditorium with the mike system and sound system is really good. The snacks provided was good.
0x3e5f5d37a6e559bd:0xc72489cc2605d63	1.045853e+20	5	1576167487	A state of the art Medical school, that offers quality health education.
0x3e5f5d37a6e559bd:0xc72489cc2605d63	1.010680e+20	5	1565072706	Excellent place for scientific and educational meeting
0x3e5f5d37a6e559bd:0xc72489cc2605d63	1.062560e+20	5	1573293200	It is an amazing place. This is where I go for university.
0x3e5f5d37a6e559bd:0xc72489cc2605d63	1.003549e+20	4	1554402027	It’s huge. Used for many conferences.there is ample parking. Basement parking available. Many conference halls. They need to improve in signage of conference halls.

Especially, for the following observations in the second dataset:

- Google ID is very complex
- Author Id desplayed as complex number, not character
- EPOCH format of data

Hence, prior to starting any analysis of the data, the dataset must be put in order.

2.3.0.3 Trends Data - Google Trends

Google Trends data was collected from 2004 to 2020 for the following keywords and regions:

- Worldwide: university in Dubai
- UAE: bachelors in Dubai
- UAE: online bachelors in Dubai

- UAE: masters in Dubai
- UAE: online masters in Dubai

Google Trends data is only avaliable in relative figures - 0 to 100 on a scale, where 100 represents the biggest interest in the keyword through Google search.

Google Trends Data

The composition of Google Trends Data is discussed in the respective section.

By default, the world data has 197 observations with 2 variables.

2.4 Steps of this project

In order to have a somewhat accurate representative opinion on education landscape in Dubai for the future, we require to analyse 3 groups of data. The project consists of the following sections for analysis of provided datasets:

Universities Census Analysis

1. We will analyze the provided data
2. We will preprocess the provided data and split the database into 2 parts: train and test with a precise explanation on the ratio
3. We will explore several methods for our models and compare RMSE & MAPE values:

- Muliple Linear Regression,
- Quantile Regression Model and
- Random Forest Model

4. We will record and analyze the results
5. We will conclude the discussion on the methods used with observed limitations in the joint results section

Google Maps Reviews Analysis

1. We will analyze the provided data
2. We will preprocess the provided data and split the database into 2 parts: train and test with a precise explanation on the ratio
3. We will explore the following method for our model and compare RMSE values:

- Factorization Model (via recosystem),
- Ensamble Model,
- Sentiment Analysis

4. We will record and analyze the results
5. We will conclude the discussion on the methods used with observed limitations in the joint results section

Google Trends Analysis

1. We will analyze the provided data
2. We will ensure that the data is provided in time-series format for our analysis
3. We will explore the following method for our model:

- Holt-Winters method (via forecast)

4. We will record and analyze the results
5. We will conclude the discussion on the methods used with observed limitations in the joint results section

3 Universities Census Analysis

3.1 Preprocessing the data

3.1.0.1 Cleaning the data

First, we would prefer to assign appropriate names for the columns in both datasets and join the datasets by respective name of the university.

Then, due to the fact, that we do not have all the required information on student numbers (drops, transfers, new enrollments), we can only consider information of total enrolled students in a particular year.

Lastly, we wil apply Label Encoder to represent the categorical data.

The summary of the updated dataset HigherEdData is presented as follows:

Table 7: Summary of HigherEdData

skim_type	skim_variable	n_missing	numeric.p0	numeric.mean	numeric.sd	numeric.p100	numeric.hist
character	Location	0	NA	NA	NA	NA	NA
character	Status	0	NA	NA	NA	NA	NA
numeric	University_Name	0	1.00000	14.413136	8.739477e+00	31.00000	
numeric	Year	0	2013.00000	2016.356992	1.992748e+00	2019.00000	
numeric	Lat	0	25.06942	25.113544	2.599420e-02	25.21348	
numeric	Long	0	55.14229	55.302732	1.280673e-01	55.41942	
numeric	Major	0	1.00000	5.471398	3.557918e+00	13.00000	
numeric	Level	0	1.00000	1.542373	4.984654e-01	2.00000	
numeric	Students	0	1.00000	186.823093	2.619611e+02	1578.00000	
numeric	UniId	0	1.00000	15.187500	9.192868e+00	34.00000	
numeric	Avarage_Tuition_Fee	0	35000.00000	64254.237288	2.989623e+04	160000.00000	
numeric	Ranking	0	2.00000	3.840042	8.706639e-01	5.00000	
numeric	Curriculum_Rating	0	1.00000	1.734110	8.967135e-01	3.00000	
numeric	YearsinDubai	0	1.00000	15.957627	8.148041e+00	25.00000	

The data looks clean to proceed with the next step.

3.2 Data Exploration

We will start the data exploration by looking at the distribution of Universities ranking. We can see that Universities with ID 11 to 30 are rated below 4 on avarage.

Distribution of students over the years also varies from university to university

The above plot provides summarized statistics by date. We can see several universities, where the data is not recorded for some of the years. Some universities increase their numbers over the years, some - decrease. The avarage is also displayed, it almost repeats the ranking avarage.

We will now explore correlation between the avaliable data variables. We will only select those variable which show correlation different from 0.

Scatterplots of each pair of numeric variable are drawn on the left part of the figure. Pearson correlation is displayed on the right. Variable distribution is available on the diagonal.

Tuition Fees are highly correlated with Ranking. The higher the avarage of the tuition fees, the better the ranking of the university. Ranking has negative correlation with Curriculum Rating (explained by difference in scales).

Years in Dubai variable appears to be insignificant, as well as it is not possible to see direct link between number of enrolled students and other variables. Major and Level of studies are also considered to have low impact on other variables.

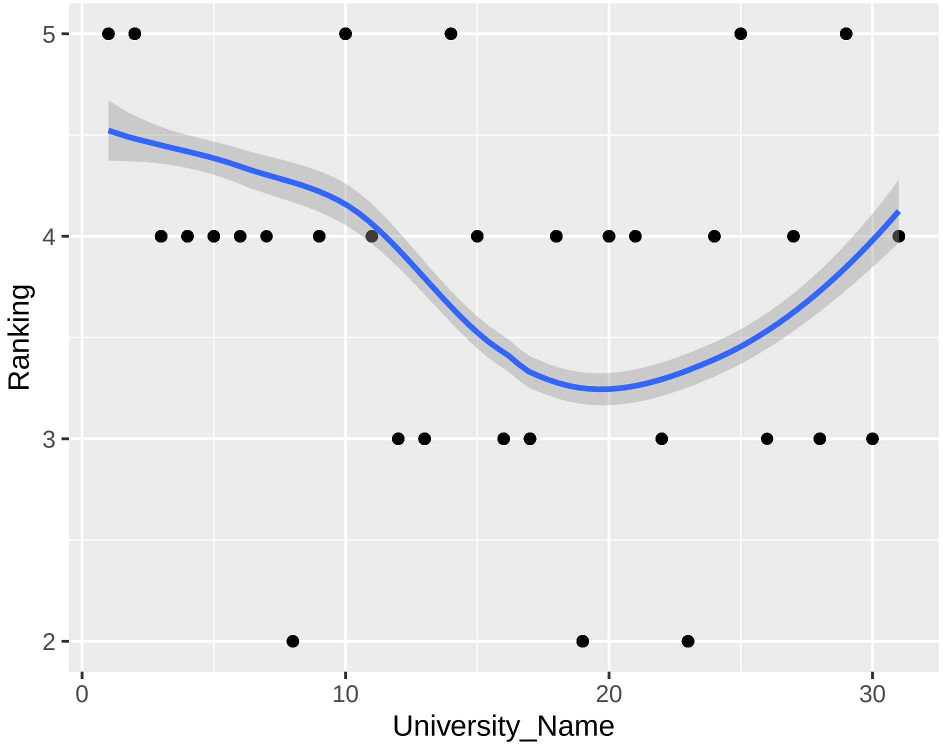


Figure 1: Distribution of Universities Rankings

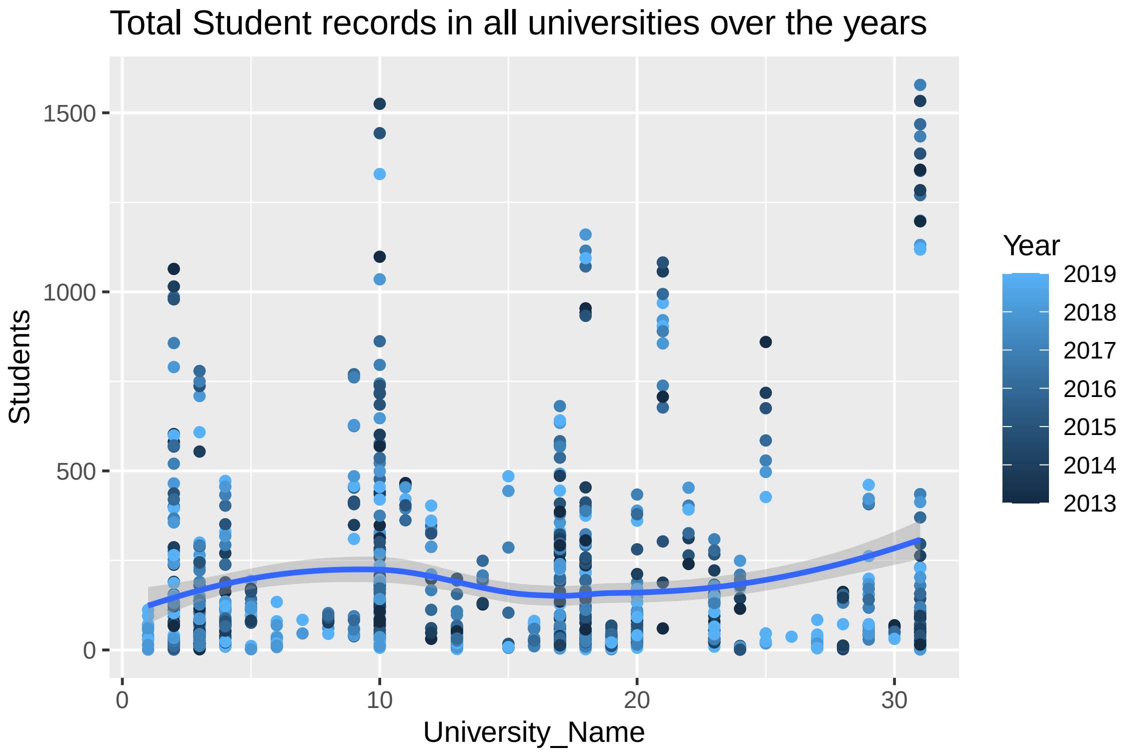


Figure 2: Distribution of Students in Universities over the years

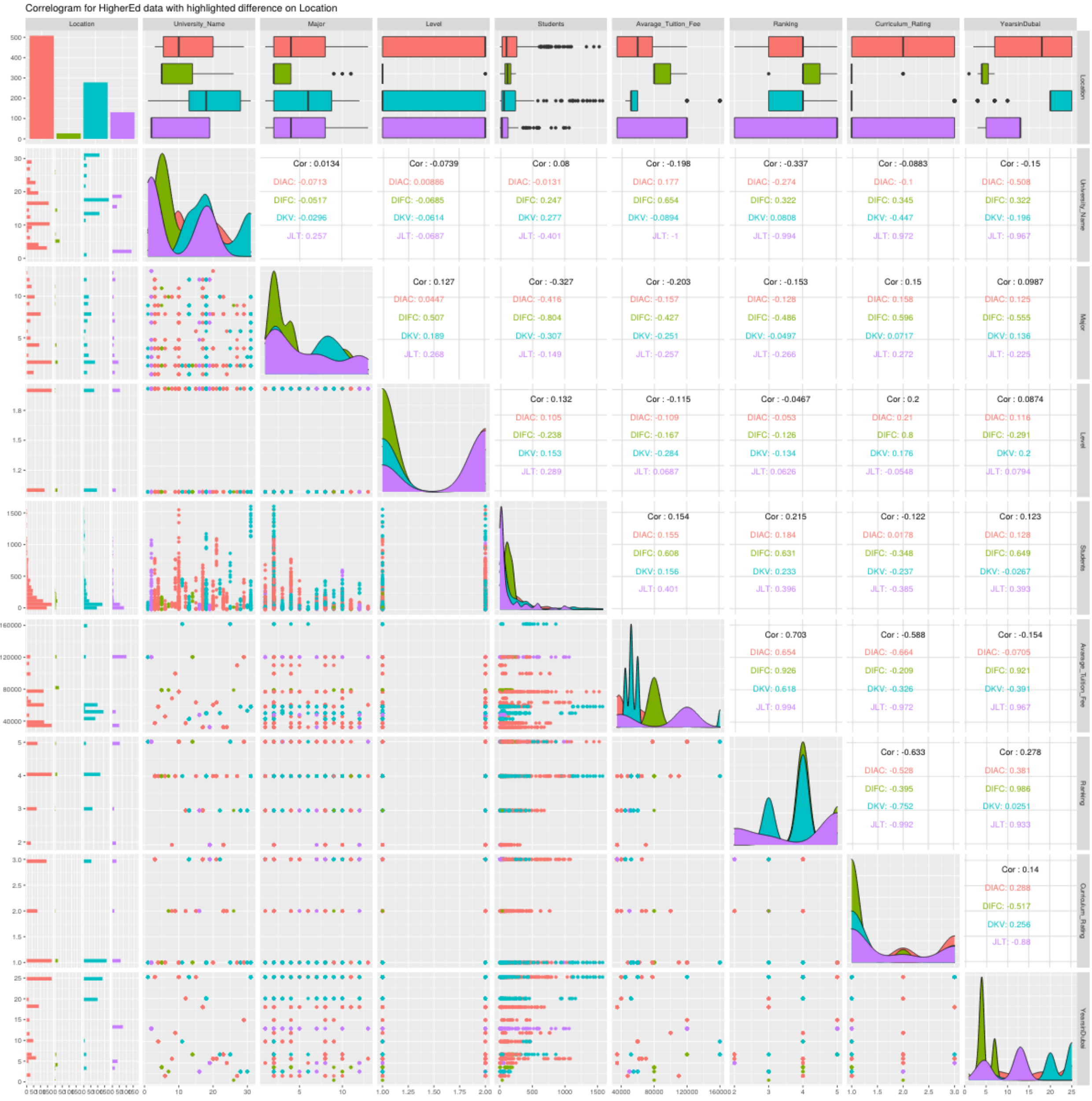


Figure 3: Correlogram for HigherEd data with highlighted difference on Location

The distributuin of location shows prevealing DIAC and very small index for DIFC. The oldest universities are located in Knowledge Village, and the youngest - in DIFC.

In general, the avaliable dataset does not provide much of an insights, as discribed in the limitations, however it might be helpful to review several models to predict the number of students in those universities.

It appears that there is no significant correlation between our variables. While for other variables it demonstrates complete autonomy, real Sessions are somewhat corelated with Duration of the sessions (+0.51). This index is still to low to conclude that they are inter-dependent. In reality, each session could be long and short, the correlation appears to be a pattern, but most likely just due to noisy data.

It will be very difficult for our model to predict the values, as there is no significant pattern detected. However, we will test each model with train and test sets and validate the final model with a portion of exsisting data to explore this.

3.3 Data Split

We will split our data into train and test sets for our linear the models. We will apply the following ratios for splitting the data:

The final model would need to make the prediction and depending on how far off it is from the actual numbers would allow us to make a choice. We will divide into 80% and 20% as general practise in Data Science and recommendation of creaters of dataPreparation package. During the production of the project, sets of 70/75/80/85/90 were tried and based on the desirable outcomes value (lower RMSE,MAPE), the selection was made for 80% split.

- Train set: 80%
- Test set: 20%

3.4 Modeling Methods

We will utilize and compare several methods/models in our research. Below overview would help us understand some theoretic background of the models.

3.4.0.1 RMSE and MAPE

In this project, we used RMSE (Root Mean Squared Error) function described in the textbook.

> In this section, we describe how the general approach to defining "best" in machine learning is to define a loss function, which can be applied to both categorical and continuous data.The most commonly used loss function is the squared loss function…The Netflix challenge used the typical error loss: they decided on a winner based on the residual mean squared error (RMSE) on a test set…

We calculate RMSE as the formula below:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n} e_t^2}$$

And MAPE (Mean absolute percentage error) as follows:

$$\text{MAPE} = \frac{100\%}{n}\sum_{t=1}^{n}\left|\frac{e_t}{y_t}\right|$$

3.4.0.2 Linear Model

This linear model take sinto account all avaliable variables and possible user biases. This model can be described mathematically as:

$$Y_{u,i} = \mu + b_i + b_u + \varepsilon_{u,i}$$

We will calculate RMSE and MAPE for this model to determine its' accuracy. We will not perform regularization for this model, as the nature of the dataset does not allow us to that. However, in our calculation of RMSE we will use log instead of actual values - to ensure the RMSE value is as small as intented.

3.4.0.3 Quantile Linear Regression Model

> Quantile regression is a type of regression analysis used in statistics and econometrics. Whereas the method of least squares estimates the conditional mean of the response variable across values of the predictor variables, quantile regression estimates the conditional median (or other quantiles) of the response variable. Quantile regression is an extension of linear regression used when the conditions of linear regression are not met.
>
> The function computes an estimate on the tau-th conditional quantile function of the response, given the covariates, as specified by the formula argument. Like lm(), the function presumes a linear specification for the quantile regression model, i.e. that the formula defines a model that is linear in parameters. For non-linear quantile regression see the package nlrq(). The function minimizes a weighted sum of absolute residuals that can be formulated as a linear programming problem. As noted above, there are three different algorithms that can be chosen depending on problem size and other characteristics. For moderate sized problems (n << 5,000, p << 20) it is recommended that the default “br” method be used. There are several choices of methods for computing confidence intervals and associated test statistics. See the documentation for summary.rq for further details and options.

3.4.0.4 Random Forest Model

Random forests are a very popular machine learning approach that addresses the shortcomings of decision trees using a clever idea. The goal is to improve prediction performance and reduce instability by averaging multiple decision trees (a forest of trees constructed with randomness). It has two features that help accomplish this. The Random Forests described in the textbook using library(randomForest). We will use the same library to perfom the operations.

3.5 Universities Census Analysis Results

3.5.1 Multiple Linear Regression Model

As discussed in methods, we would like to account for all possible biases in the student numbers and see how accurate our prediction could be.

The naive avarage for our dataset is

```
## [1] "Naive avarage: 186.973509933775"
```

```
## [1] "Avarage with log1p: 5.23630104829438"
```

After applying linear model function lm() we have received the following results:

```
# Applying Muliple Linear Regression
regressor = lm(Students ~ Year+University_Name+Major+Level+
                 Ranking+Avarage_Tuition_Fee+Curriculum_Rating+
                 YearsinDubai,data = train)

# Cearting the input test dataframe for testing the model for the prediction.
x_test = test[, c("Year","University_Name","Major","Level",
                  "Avarage_Tuition_Fee","Ranking","Curriculum_Rating","YearsinDubai")]

# Actual output of the Testing column
y_test = test$"Students"

# Applying the prediction of the regression model on the test dataframe.
y_pred = predict(regressor, newdata = x_test)

# Getting the absolute if there is any negitive value.
y_pred = abs(y_pred)

# Combining the Predicted Values and Actual Values into one dataframe.
actuals_preds <- data.frame(cbind(actuals=y_test, predicteds=y_pred))

# Computing the RMSE.
rmse <- sqrt(mean(as.matrix(log1p(y_test) - log1p(y_pred))^2))
print(paste('RMSE:', rmse))
```

```
## [1] "RMSE: 1.60048202904415"
```

```
# MeanAbsolutePercentageError (MAPE)
mape <- mean(abs((y_pred - y_test))/y_test)
print(paste('MAPE:',mape))
```

```
## [1] "MAPE: 8.55772934842259"
```

We can see that RMSE value is quite high even after adjusting the values as matrix and log.

We shall try to optimize these values with the next model.

3.5.2 Quantile Linear Regression Model

Thq Quantile Linear Regression Model uses τ value = 0.25. We received the following results:

```
## [1] "RMSE: 1.57035206359146"
```

```
## [1] "MAPE: 2.00218653204983"
```

```
##     actuals          predicteds
##  Min.   :   1.0   Min.   : 3.154
##  1st Qu.:  24.0   1st Qu.:29.335
##  Median :  79.0   Median :39.917
##  Mean   : 186.2   Mean   :40.852
##  3rd Qu.: 258.0   3rd Qu.:53.683
##  Max.   :1329.0   Max.   :84.937
```

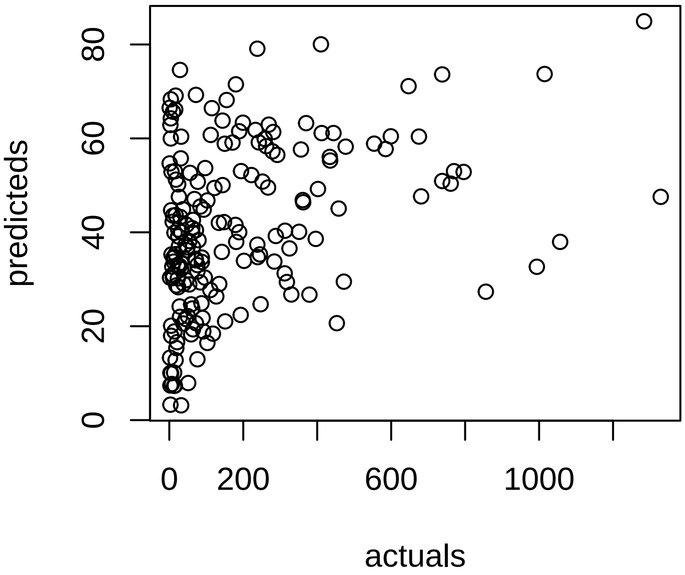


Figure 4: Predict vs Actuals for Quantile Linear Reg Model Plot

This significantly improved our MAPE value, however the RMSE has increased. We shall explore other model to try achieve better results.

3.5.3 Random Forest Model

Now we will try to predict amount of students, total, over the years. this decision tree appears to be more complex than for other parameteres, however it can quite accurately predict amount of students enrolled at the university based on other factors. Interestingly, 27% of students can be determined based on choice of their Major and Level of studies. It means that this model does not account for these 27% as their sort level is very small. If the avarage Tuition Fees is less than 46e+3 and the University ID is less than 21, The rating i slikely be 3, as 28% of the universities data.

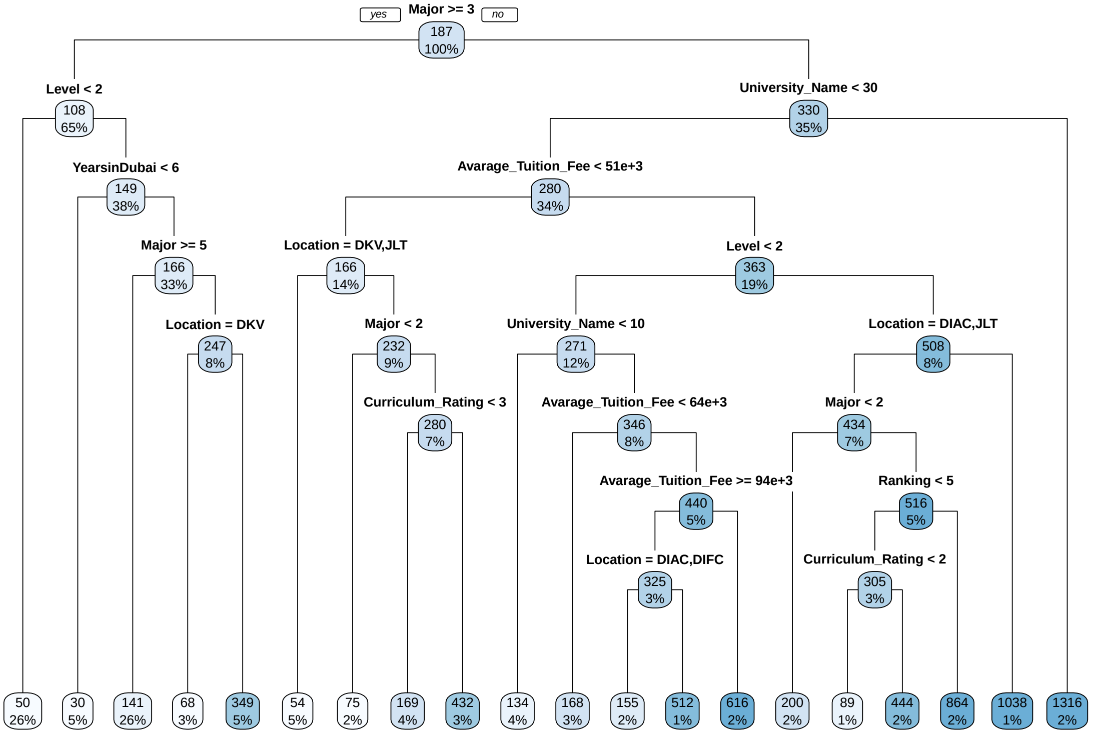


Figure 5: Deicsion Tree for Student Numbers

We will now apply RandomForest function for Students using other avaliable parameters.

```
## [1] "RMSE: 1.19336616873873"
```

```
## [1] "MAPE 4.52983274254009"
```

Both, RMSE and MAPE values have decreased. Random Forest, perhaps, is the best fit for the provided data. We will use RandomForestExplainer package to get to the bottom of our model.

Distribution of minimal depth

The plot below shows the distribution of minimal depth among the trees of our forest. Note that:

- the mean of the distribution is marked by a vertical bar with a value label on it (the scale for it is different than for the rest of the plot),

- the scale of the X axis goes from zero to the maximum number of trees in which any variable was used for splitting.

Minimal depth for a variable in a tree equals to the depth of the node which splits on that variable and is the closest to the root of the tree. If it is low than a lot of observations are divided into groups on the basis of this variable.

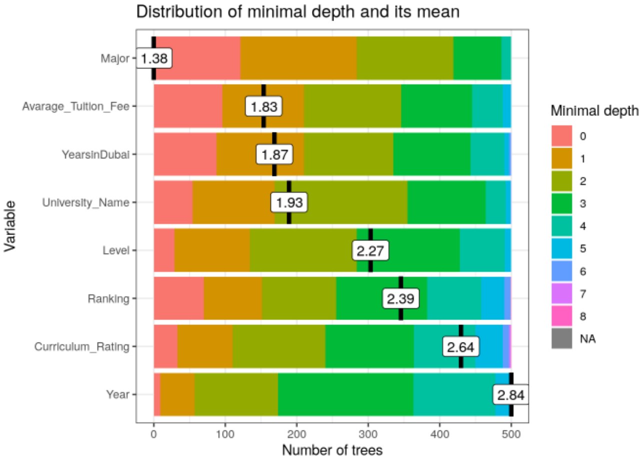


Figure 6: Minimal Depth Distribution

Multi-way importance plot

The multi-way importance plot shows the relation between three measures of importance and labels 10 variables which scored best when it comes to these three measures (i.e. for which the sum of the ranks for those measures is the lowest). The first multi-way importance plot focuses on three importance measures that derive from the structure of trees in the forest:

- mean depth of first split on the variable,
- number of trees in which the root is split on the variable,
- the total number of nodes in the forest that split on that variable.

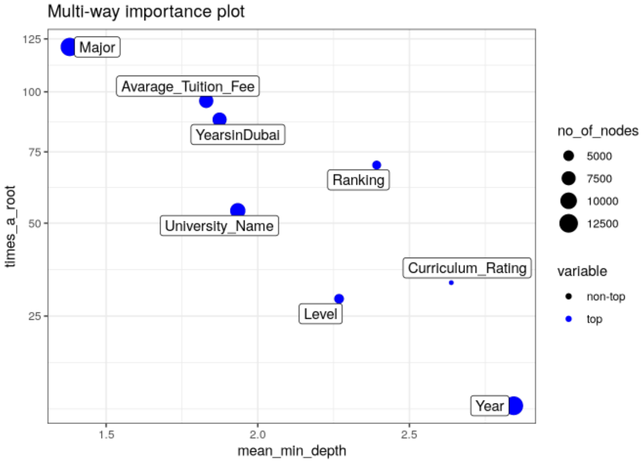


Figure 7: First Multi-way importance plot

The second multi-way importance plot shows importance measures that derive from the role a variable plays in prediction: with the additional information on the p-value based on a binomial distribution of the number of nodes split on

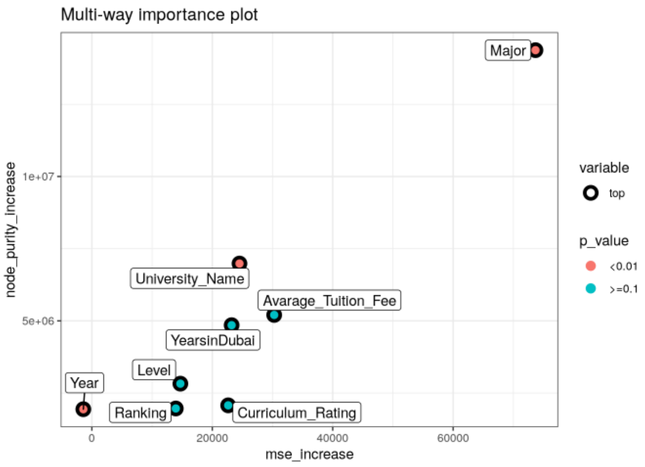


Figure 8: Second Multi-way importance plot

the variable assuming that variables are randomly drawn to form splits (i.e. if a variable is significant it means that the variable is used for splitting more often than would be the case if the selection was random).

Comparing rankings of variables

The plot below shows bilateral relations between the rankings of variables according to chosen importance measures. This approach might be useful as rankings are more evenly spread than corresponding importance measures. This may also more clearly show where the different measures of importance disagree or agree.

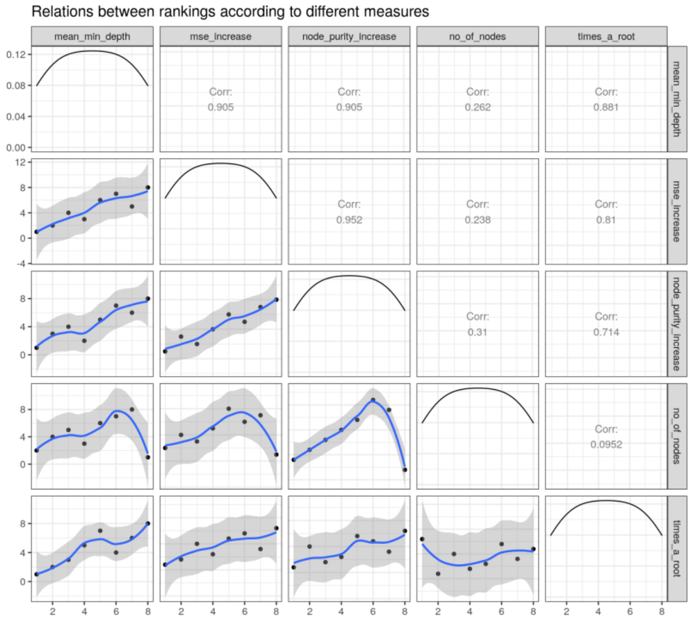


Figure 9: Bilateral relations between the rankings of variable

Variable interactions

Conditional minimal depth

The plot below reports 30 top interactions according to mean of conditional minimal depth – a generalization of minimal depth that measures the depth of the second variable in a tree of which the first variable is a root (a subtree of a tree from the forest). In order to be comparable to normal minimal depth 1 is subtracted so that 0 is the minimum. For example value of 0 for interaction x:y in a tree means that if we take the highest subtree with the root splitting on x then y is used for splitting immediately after x (minimal depth of x in this subtree is 1). The values presented are means over all trees in the forest.

Note that: - the plot shows only 30 interactions that appeared most frequently, - the horizontal line shows the minimal value of the depicted statistic among interactions for which it was calculated, - the interactions considered are ones with the following variables as first (root variables): and all possible values of the second variable.

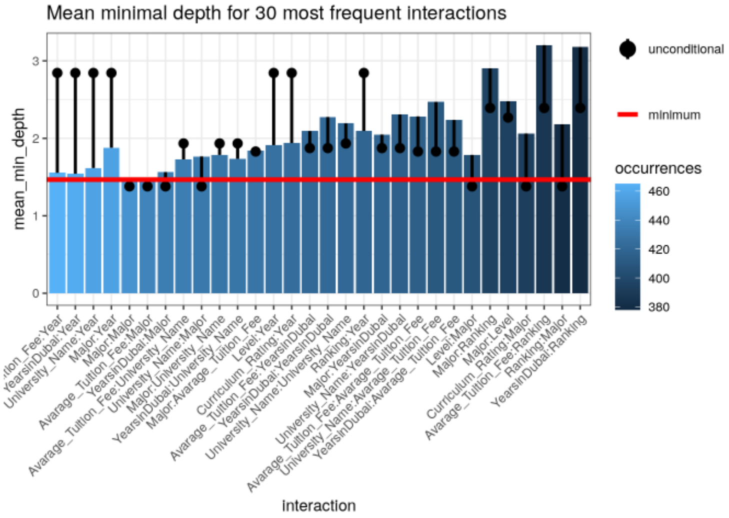


Figure 10: Variable Interactions - Conditional Minimal Depth

More detailed explanation of the above graphs is also avaliable in a complied report.

4 Google Map Reviews Analysis

4.1 Preprocessing the data

4.1.0.1 Cleaning the data

As this analysis includes Factorization model, Ensambles as well as Sentiment Analysis, the data needs to be prepared in accordance with each method.

However, there are few common steps with data manipulation which we need to perform in order to continue with the analysis.

From initial data summary, we noticed that GUni and GReviews have common field of google_id which is very complex and might cause issues if not converted.

Secondly, we need to get rid of columns which we will not utilize in our report.

We noticed, that after joining the tables through inner_join function, some values have dissapeared - it is simply due to the fact that in the rating dataset we have ratings for organizations, which are not considered to be universities, which we cleaned before importing the data.

Table 8: First six records of GF

unid	userid	review_rating	review_timestamp	name
1	1	4	2018-12-23 07:52:28	Cass Business School Dubai International Financial Centre
2	2	5	2019-08-03 07:59:19	Mohammed Bin Rashid University Of Medicine and Health Sciences
2	3	5	2019-12-12 16:18:07	Mohammed Bin Rashid University Of Medicine and Health Sciences
2	4	5	2019-08-06 06:25:06	Mohammed Bin Rashid University Of Medicine and Health Sciences
2	5	5	2019-11-09 09:53:20	Mohammed Bin Rashid University Of Medicine and Health Sciences
2	6	4	2019-04-04 18:20:27	Mohammed Bin Rashid University Of Medicine and Health Sciences

We can now look at the summary of the updated data.

Table 9: Summary of GF

skim_type	skim_variable	n_missing	numeric.p0	numeric.mean	numeric.sd	numeric.p100	numeric.hist
character	name	0	NA	NA	NA	NA	NA
numeric	unid	0	1	37.431396	23.265417	76	▇▆▅▆▆
numeric	userid	0	1	1519.510770	921.499428	3161	▇▇▇▇▇
numeric	review_rating	0	1	4.295958	1.258754	5	▁▁▁▂▇
POSIXct	review_timestamp	0	NA	NA	NA	NA	NA

4.2 Data Exploration

We will now explore our joined dataset GF.

It appears, there are more 5 star-reviews on google data.

The distribution of ratings by univerity also show the follwoing data:

The number of ratings also not fistributed equally and varies from university to university.

Majority of users left only one rating.

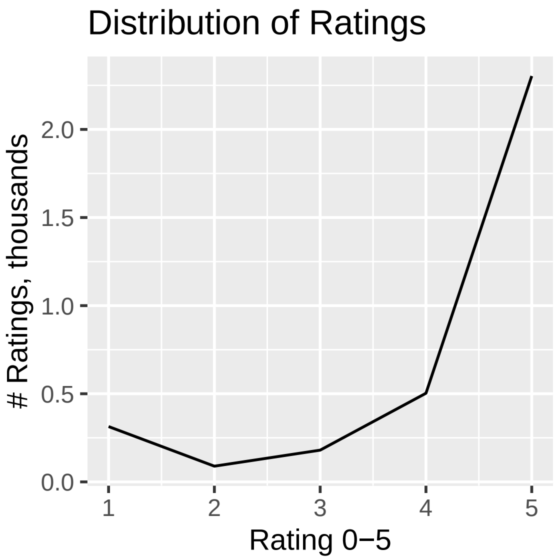


Figure 11: Distribution of ratings

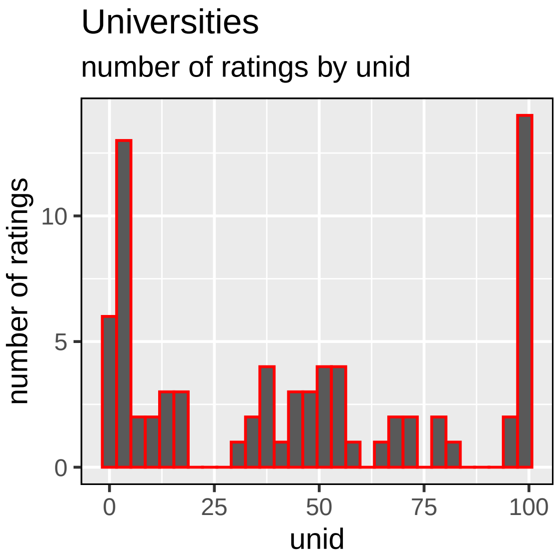


Figure 12: Distribution of ratings

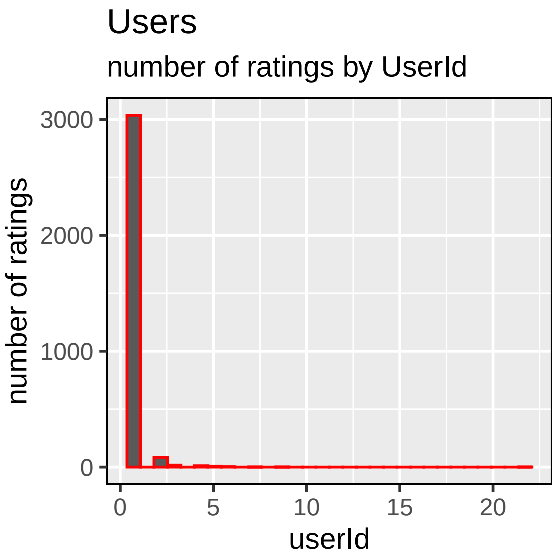


Figure 13: Distribution of ratings

There are few who rated several organizations - could be the students who transfered from university to university or continued postgraduate studies.

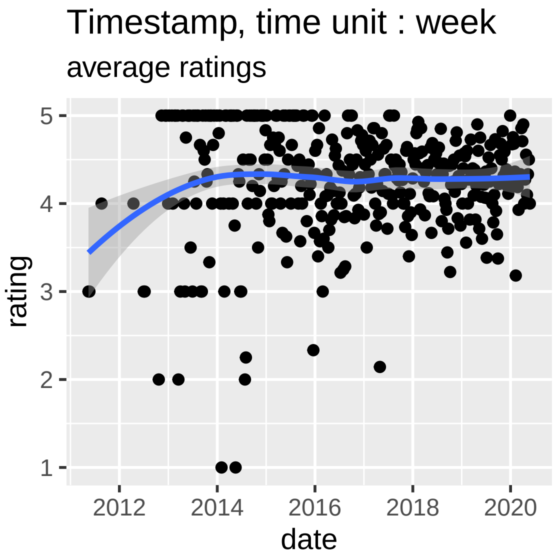


Figure 14: Distribution of ratings over years

In the past years, we can notice the distribution has shfted from 5 to somewhat above 4. 2013-2016 were years where the avarage was the highest. It might be possible to explain this observation as in the recent years, there is less interest/satisfaction in local universities. We cannot determine this as the data does not provide detailed infomration on those people who have rated the universities.

Correlogram showcases how different variables are (not) correlated with each other as well as the distribution over the years. In the dates we can notice seasonality - perhpaps the time of fininshing the Fall vs Spring sessions.

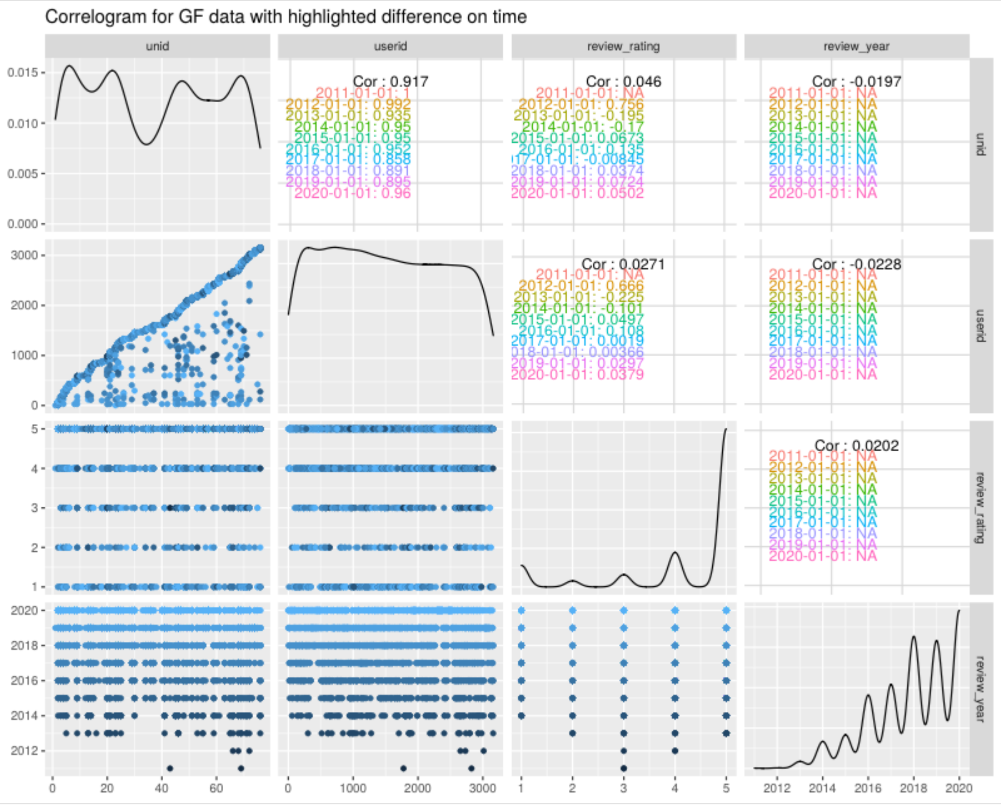


Figure 15: Correlogram for GF data with highlighted difference on time

4.3 Data Split

We will split our data into train and test sets. We will apply the following ratios for splitting the data:

The final model would need to make the prediction and depending on how far off it is from the actual numbers would allow us to make a choice. We will divide into 80% and 20% as general practise in Data Science and recommendation of creaters of dataPreparation package.

- Train set: 80%
- Test set: 20%

As confirmed in the EdX forum, for this project we will not be using validation set.

We will use dataPreparation package to help us split the data into 80/20 ratio and check for inconsistancies in the data:

```
## Warning in set.seed(1, sample.kind = "Rounding"): non-uniform 'Rounding' sampler
## used
```

```
## [1] "whichAreConstant: it took me 0.01s to identify 0 constant column(s)"
```

```
## [1] "whichAreInDouble: it took me 0s to identify 0 column(s) to drop."
```

```
## [1] "whichAreBijection: unid is a bijection of name. I put it in drop list."
## [1] "whichAreBijection: it took me 0.02s to identify 1 column(s) to drop."
```

There are no constant, double columns. The bijection column name was removed from our dataset.

4.3.0.1 Scaling

Most machine learning algorithm rather handle scaled data instead of unscaled data. To perform scaling (meaning setting mean to 0 and standard deviation to 1), function fastScale is available. Since it is highly recommended to apply same scaling on train and test, we will compute the scales first using the function build_scales. However, after cleaning the dataset there is no need for addditional scaling.

4.3.0.2 Discretization

We do not have variables to perform discretization on.

4.3.0.3 Encoding Categorical

We do not have categorical data for this dataset.

4.3.0.4 Controling shape

Last, but not least, it is very important to make sure that train and test sets have the same shape (for example the same columns).

To make sure of that we will perform the following function sameShape:

```
## [1] "sameShape: verify that every column is present."
## [1] "sameShape: drop unwanted columns."
## [1] "sameShape: verify that every column is in the right type."
## [1] "sameShape: verify that every factor as the right number of levels."
## [1] FALSE
## [1] FALSE
## [1] FALSE
## [1] FALSE
## [1] FALSE
```

SameShape reported that our data is in order.

Now we are in possessions of 2 datasets as follows:

Feature	Train	Test
Number of Rows	2711	678
Number of Columns	5	5
Unique Universities	75	65
Unique Users	2551	663
Max rating	5	5

4.3.1 Data for Sentiment Analysis

As Sentiment Analysis requires diferent set of avaliable data, we shall adjust our data accordingly. We will utilize similar steps as the previous analysis. We will merge data from GFUni and GFReview to a newly created GFS dataset. We will assign proper IDs for universities and users. We will change the EPOCH timestamp into readable format. We should also filter all empty reviews.

Table 10: First six records of GFS

unid	userid	review_rating	review_timestamp	review_text
1	1	4	2018-12-23 07:52:28	One of the top-ranked universities in the world and in the region. Cass Business School is the ultimate place to receive an executive MBA. The weekend modular format of the program makes it easy to balance between work, life and study.
2	2	5	2019-08-03 07:59:19	Attended a CME by Cardiff University and Emirates Diabetes Society conducted at Mohammed Bin Rashid University of Medicine and health Sciences. The venue is excellent. The auditorium with the mike system and sound system is really good. The snacks provided was good.
2	3	5	2019-12-12 16:18:07	A state of the art Medical school, that offers quality health education.
2	4	5	2019-08-06 06:25:06	Excellent place for scientific and educational meeting
2	5	5	2019-11-09 09:53:20	It is an amazing place. This is where I go for university.
2	6	4	2019-04-04 18:20:27	It's huge. Used for many conferences.there is ample parking. Basement parking available. Many conference halls. They need to improve in signage of conference halls.

We can now look at the summary of the updated data.

Table 11: Summary of GFS

skim_type	skim_variable	character.min	character.max	character.empty	character.n_unique	character.whitespace	numeric.hist
character	review_text	1	2683	0	1579	6	NA
factor	review_rating	NA	NA	NA	NA	NA	NA
numeric	unid	NA	NA	NA	NA	NA	▆▆▅▇▅
numeric	userid	NA	NA	NA	NA	NA	▆▇▆▇▆
POSIXct	review_timestamp	NA	NA	NA	NA	NA	NA

4.3.1.1 Data Cleaning

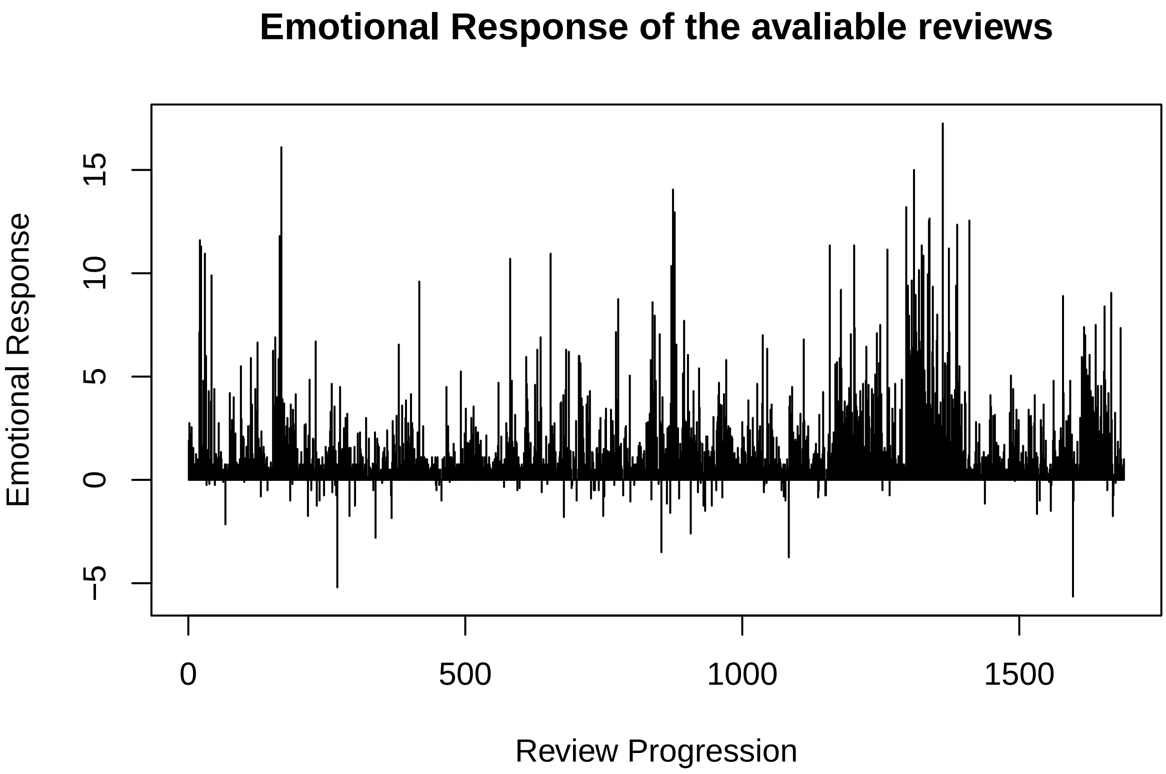
We are ready to start preprocessing the 1689 reviews that contain text. We should further clean the data: * removing numbers * removing stopwords (words that don't have any contextual meaning e.g. "and", "so" etc.) * Assigning to each rating a sentiment score via syuzhet library

For our combined score we will use several lexicon libraries, sich as bing, afinn and nrc.

The sumary of originated vector syuzhet:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  -5.650   0.500   1.100   1.791   2.500  17.250
```

The total score distribution is now ready for our analysis:



4.4 Modeling Methods

4.4.0.1 Matrix Factorization

From the available literature review, we can gather that a comprehensive analysis was made to evaluate various methods for training the algorithms for recommendations. Specifically, we are referring to practical comparisons made by Taras Hnot, where he argues that:

> There are two main categories of collaborative filtering algorithms: memory-based and model-based methods.
>
> Memory-based methods simply memorize all ratings and make recommendations based on relation between user-item and rest of the matrix. In model-based methods predicting parametrized model firstly is needed to be fit based on rating matrix and then recommendations are issued based on fitted model.
>
> Model-based methods, on the other hand, build parametrized models and recommend items with the highest rank, returned by model.

In his research, Taras has pointed out that "the best performance was shown by Matrix Factorization techniques with Stochastic Gradient Descend".

In this project we will utilize recosystem, as it is intuitively more clear and uses simple syntax.

recosystem is an R wrapper of the LIBMF library developed by Yu-Chin Juan, Yong Zhuang, Wei-Sheng Chin and Chih-Jen Lin, an open source library for recommender system using marix factorization.

> The main task of recommender system is to predict unknown entries in the rating matrix based on observed values. Each cell with number in it is the rating given by some user on a specific item, while those marked with question marks are unknown ratings that need to be predicted. In some other literatures, this problem may be given other names, e.g. collaborative filtering, matrix completion, matrix recovery, etc. Highlights of LIBMF and recosystem LIBMF itself is a parallelized library, meaning that users can take advantage of multicore CPUs to speed up the computation. It also utilizes some advanced CPU features to further improve the performance. [@LIBMF] recosystem is a wrapper of LIBMF, hence the features of LIBMF are all included in recosystem. Also, unlike most other R packages for statistical modeling which store the whole dataset and model object in memory, LIBMF (and hence recosystem) is much hard-disk-based, for instance the constructed model which contains information for prediction can be stored in the hard disk, and prediction result can also be directly written into a file rather than kept in memory. That is to say, recosystem will have a comparatively small memory usage.

After examining the RStudio packages as an alternative to our previous model, the most intuitive and effortless solution was named as recosystem - a specified library which required minimum efforts to be executed.

4.4.0.2 Ensambles

After examining the results of recosystem, the decision was made not to utilize Ensables for our data due to limitation of variability for the choice of the rating for the university. The methods might be useless as only few users have rated more than one university.

4.4.0.3 WordCloud

We will use wordcloud to analysie sentiments of users in our dataset.

4.5 Google Map Reviews Analysis Results

4.5.1 Matrix Factorization

We followed the steps described in the package documentation, since it is a small dataset, the model run very fast.

```
## Warning in set.seed(123, sample.kind = "Rounding"): non-uniform 'Rounding'
## sampler used
```

```
## iter      tr_rmse          obj
##    0       2.9421   2.6042e+04
##    1       1.0504   5.2939e+03
##    2       0.6256   3.4026e+03
##    3       0.3907   2.7362e+03
##    4       0.3227   2.5864e+03
##    5       0.2889   2.5119e+03
##    6       0.2694   2.4663e+03
##    7       0.2558   2.4323e+03
##    8       0.2474   2.4071e+03
##    9       0.2398   2.3841e+03
##   10       0.2340   2.3660e+03
##   11       0.2261   2.3436e+03
##   12       0.2225   2.3304e+03
##   13       0.2161   2.3157e+03
##   14       0.2098   2.2999e+03
##   15       0.2051   2.2890e+03
##   16       0.1997   2.2774e+03
##   17       0.1919   2.2620e+03
##   18       0.1882   2.2553e+03
##   19       0.1818   2.2439e+03
```

After running the recosystem package on our trainSet, we can see the RMSE values of **0.1878**:

First 10 predicted values include:

```
##  [1] 4.31870 4.31870 4.31870 4.31870 4.31870 4.31870 4.31870 5.01707 4.31870
## [10] 4.31870
```

```
## [1] "RMSE: 0.0915184914186641"
```

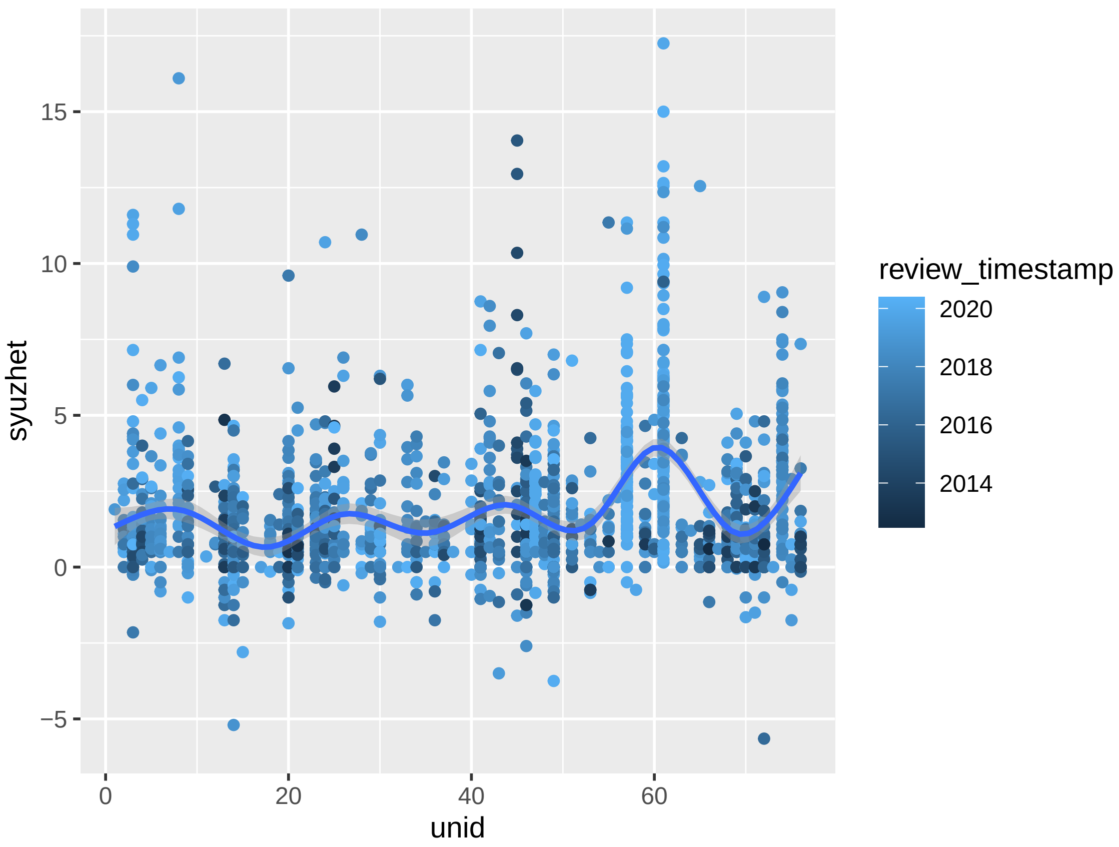
```
## [1] "MAPE: 0.528683433426048"
```

This is very good result, however we can notice non-integer values in the predicted values. Since the data is very spearse, it was expected, that the model will not work as efficient.

4.5.2 Sentiment Analysis

We will now combine emotional response scores with avaliable responses and present initial analysis for the same.

The below plot demonstrates how each university is rated according to summarized sentiment score, over the years.



Let’s consider words in the dataset which are used more than 50 times, to understand the buzz-effect.

As anticipated, there are more possitive attributes to the wordcloud of the reviews.

At the same time, it is important to note few buzz-words which universitied can take into account for their business development efforts: helpful, students, experience, quality, place, professors, people, money.

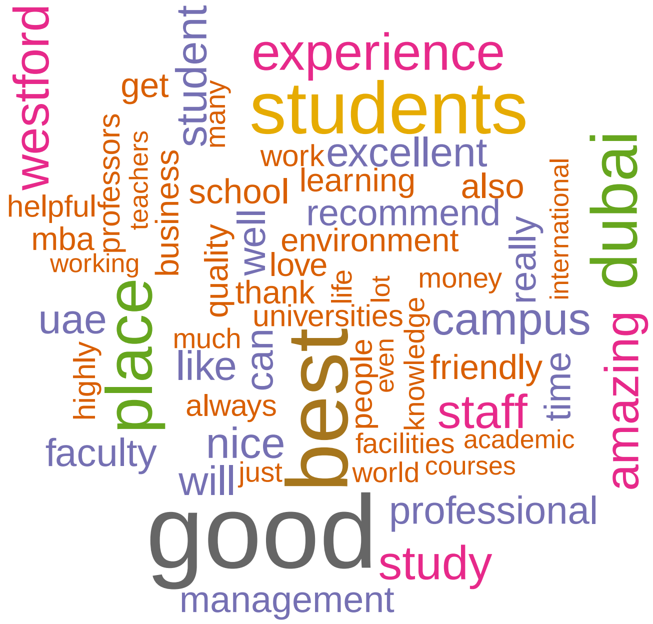


Figure 16: Wordcloud for user reviews of universities in the UAE

5 Google Trends Forecasting

5.1 Overview and Modeling Methods

5.1.0.1 Google Trends Overview

The library gtrendR allows us to collect the data from Google Trends server directly. The data is collected from 2004 until date and is relative to within the category of the keyword search (i.e. 100 for keyword A means the peak of interest in search hits for keyword A itself, not the volume of performed searches).

The package has a small limitation - if we are to inquire several keywords at the same time, the library will provide information on keywords relevant to each other. This might provide addition trouble in understanding our data, hence we will import each keyword separately and join all the data in one table google.

Here we will jump directly to the summary of provided data.

Table 12: Summary of google.trends for University in Dubai, Worldwide

skim_type	skim_variable	n_missing	numeric.mean	numeric.sd	numeric.p0	numeric.p100	numeric.hist
numeric	University_world	0	60.218274	13.616394	21	100	▁▅▇▆▁
numeric	Bachelors_world	0	17.131980	14.944893	0	100	▇▃▁▁▁
numeric	Masters_world	0	24.644670	11.573881	0	100	▂▇▁▁▁
numeric	Bachelors_UAE	0	3.822335	8.868306	0	100	▇▁▁▁▁
numeric	Masters_UAE	0	9.172589	11.669368	0	100	▇▁▁▁▁
numeric	Online_University_world	0	14.862944	12.568039	0	100	▇▂▁▁▁
numeric	Online_Masters_world	0	6.944162	9.486937	0	100	▇▁▁▁▁
POSIXct	date	0	NA	NA	NA	NA	NA

Few important points to note:

- **University_world** stands for keyword University in Dubai searched worldwide

- **Bachelors_world** stands for keyword Bachelors in Dubai searched worldwide
- **Masters_world** stands for keyword Masters in Dubai searched worldwide
- **Bachelors_UAE** stands for keyword Bachelors in Dubai searched in the UAE
- **Masters_UAE** stands for keyword Masters in Dubai searched in the UAE
- **Online_University_world** stands for keyword Online University in Dubai searched worldwide
- **Online_Masters_world** stands for keyword Online Masters in Dubai searched worldwide
- There is no data avaliable for keyword search **Online bachelors in Dubai** both in the UAE and worldwide

A more detailed view on the trends is presented below:

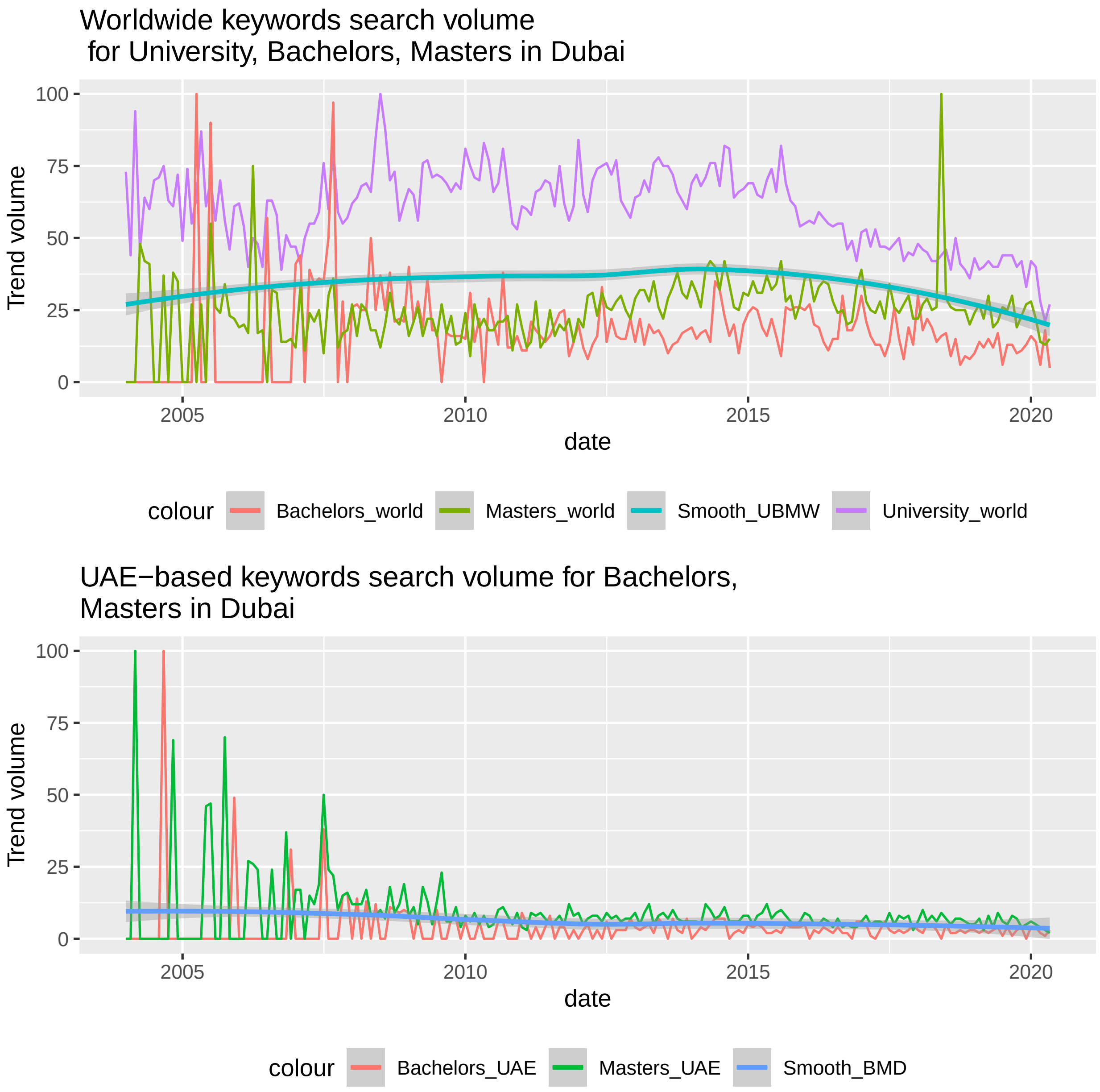


Figure 17: Worldwide Trends for studies in Dubai

Worldwide keyword search volume presents a rather declining trend for 2015-2020 with seasonal spikes and ubnormal one-time spike for masters in Dubai in 2018. The smooth line presents avarage trend for those 3 combined searches and also declines for the past 5 years.

In UAE, however, the interest for studies in Dubai appears to be more or less constant with a ligh decline from 2012. Within 15 years, the overall interest has dropped over 75%.

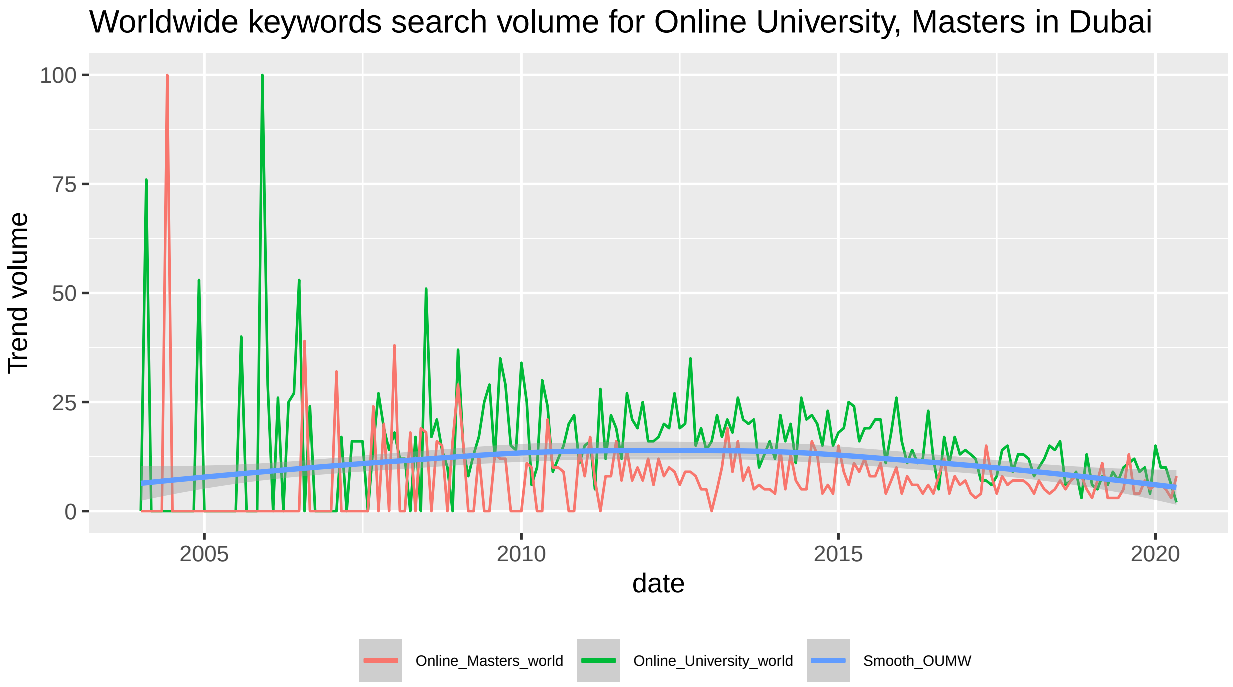


Figure 18: Worldwide Trends for studies in Dubai

Online studies in Dubai searched worldwide have been increasing frin 2010 to 2015, however declined thereafter. The overall interest has dropped from 2004-2005 values.

We will now present correlation analysis for those values to understand if there is any connection between the interest rates. We will apply discretization of the year parameter to show the data for equal 4 years period, 5 periods as follows:

```
## [1] "fastDiscretization: I will discretize 1 numeric columns using, bins."
## [1] "fastDiscretization: it took me: 0s to transform 1 numeric columns into, binarised columns."
```

```
##
## [2004, 2008[ [2008, 2012[ [2012, 2016[ [2016, 2020[ [2020, +Inf[
##           48           48           48           48            5
```

We are now ready to perform correlation analysis.

There is no significant correlation between the variables, apart from keyword searches for Bachelor studies in Dubai between worldwide and UAE-based numbers (0.4). This means our forecasts should be performed separetely for each keyword group.

5.1.0.2 Forecasting with Holt-Winters

The R package forecast provides methods and tools for displaying and analysing univariate time series forecasts including exponential smoothing via state space models and automatic ARIMA modelling.

The additive Holt-Winters prediction function (for time series with period length p) is

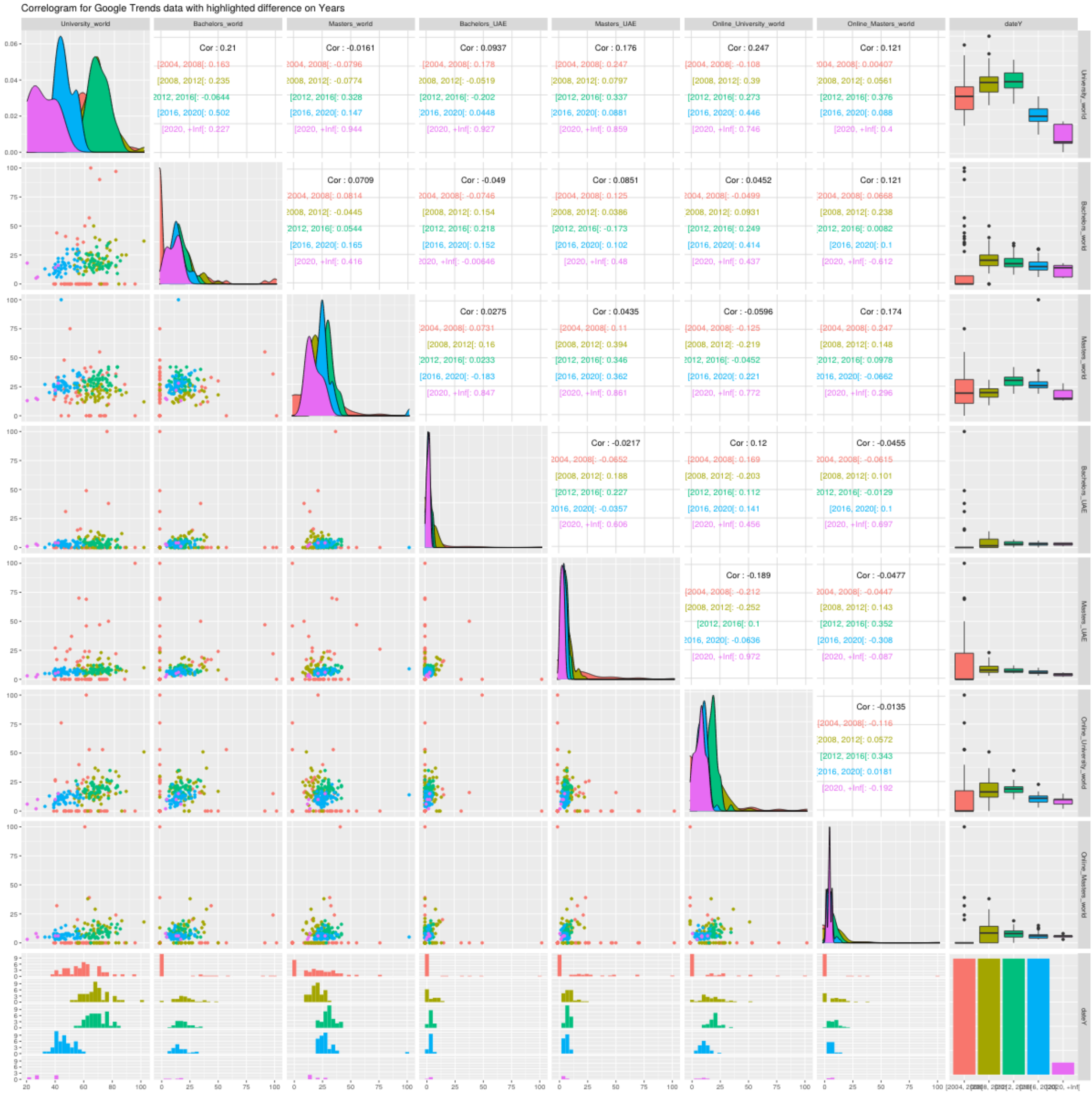


Figure 19: Correlogram for Google Trends data with highlighted difference by Year interval

$$\hat{Y}[t+h] = a[t] + hb[t] + s[t-p+1+(h-1) \text{ mod } p],$$

where $a[t]$, $b[t]$ and $s[t]$ are given by

$$\begin{aligned} a[t] &= \alpha(Y[t] - s[t-p]) + (1-\alpha)(a[t-1] + b[t-1]) \\ b[t] &= \beta(a[t] - a[t-1]) + (1-\beta)b[t-1] \\ s[t] &= \gamma(Y[t] - a[t]) + (1-\gamma)s[t-p] \end{aligned}$$

The data in x are required to be non-zero for a multiplicative model, but it makes most sense if they are all positive.

The function tries to find the optimal values of α and/or β and/or γ by minimizing the squared one-step prediction error if they are NULL (the default). optimize will be used for the single-parameter case, and optim otherwise.

We will utilize this model to predict the future trend based on the dataset.

We will check RMSE and MAPA values for one of the variables first on the train and test sets, and then apply the model for the entire data.

We divided train and set data as interval between 2004-2009 for trainset and 2010 for testset.

```
##                    ME     RMSE      MAE       MPE     MAPE      MASE
## Training set 1.599698 15.58445 11.19075 0.3253559 17.58937 0.8138726
## Test set     4.749422 14.17841 11.49176 3.1814088 18.03875 0.8357642
##                     ACF1 Theil's U
## Training set -0.07361522        NA
## Test set     -0.51649378 0.6160327
```

Our RMSE and MAPE values have very close values in our training and test sets. In the results section we will highlight the confidence intervals in our forecasts.

5.2 Google Trends Analysis Results

5.2.1 Forecasting with Holt-Winters

As we have noticed in the discovery analysis, the trend for all searches was growing significantly for first years from 2004, but from 2015 have been reducing.

We will utilize forecast library build specifically to work with time-series. We will instruct the algorithm to perform forecasts for trend analysis for the next 2 years and will take into all avaliable data from 2004 onwards.

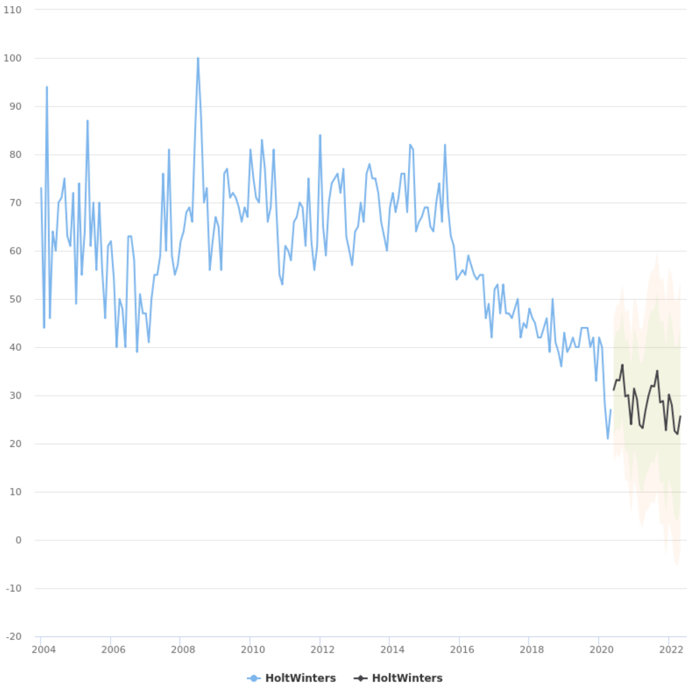


Figure 20: HoltWinters Forecast for University in Dubai, worldwide

The forecast demonstrates further significant decrease of the overall volume of searches for universities in Dubai, worldwide.

The forecast demonstrates stable interest for bachelors degrees in Dubai, worldwide.

The forecast demonstrates further decrease of the overall volume of searches for masters degree in Dubai, worldwide with seasonal spikes.

The forecast demonstrates further decrease of the overall volume of searches for masters degree in Dubai, in the UAE, with minimum seasonality.

The forecast demonstrates further insignificant decrease of the overall volume of searches for masters degree in Dubai, in the UAE, with modest seasonality.

The forecast demonstrates stable interest for online universities in Dubai, worldwide.

The forecast demonstrates significant increase of the overall volume of searches for online masters degree in Dubai, worldwide.

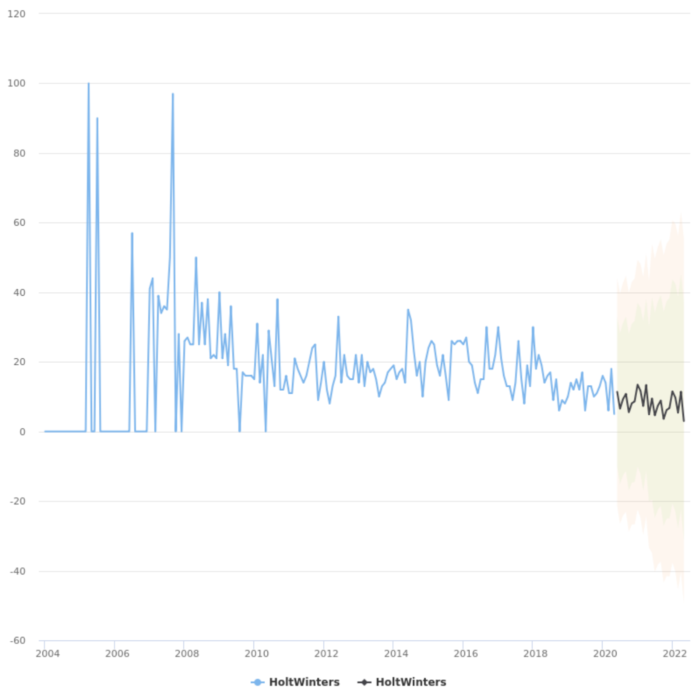


Figure 21: HoltWinters Forecast for University in Dubai, worldwide

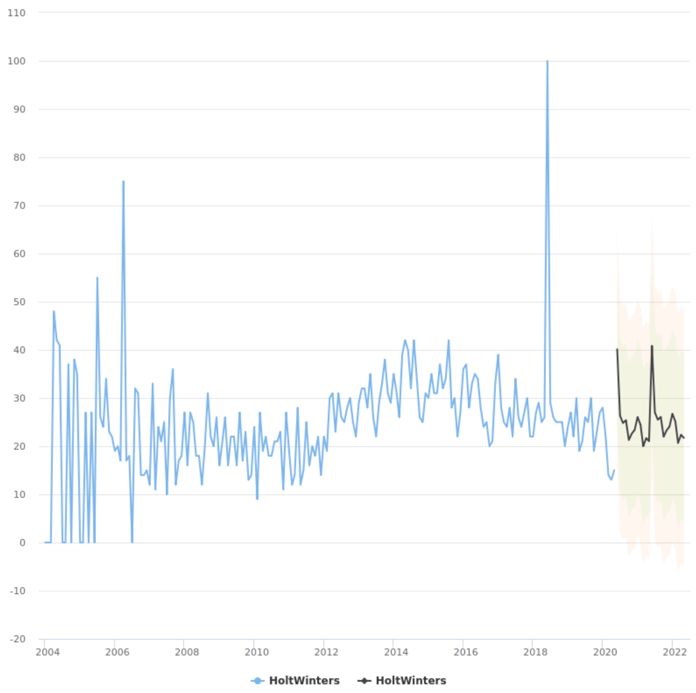


Figure 22: HoltWinters Forecast for University in Dubai, worldwide

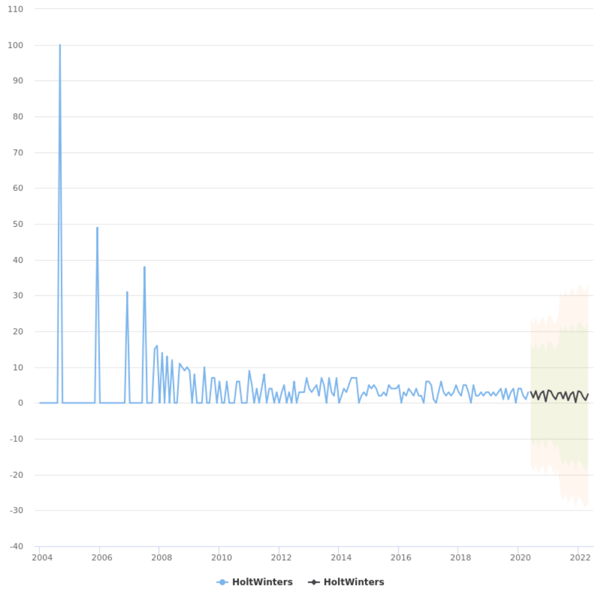


Figure 23: HoltWinters Forecast for University in Dubai, worldwide

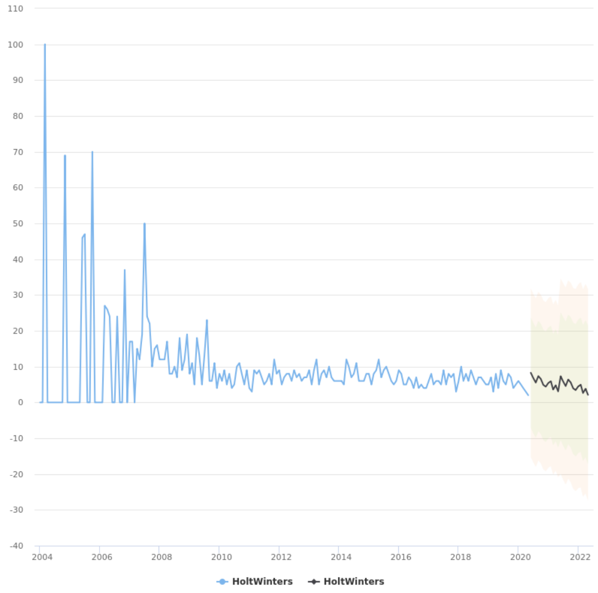


Figure 24: HoltWinters Forecast for University in Dubai, worldwide

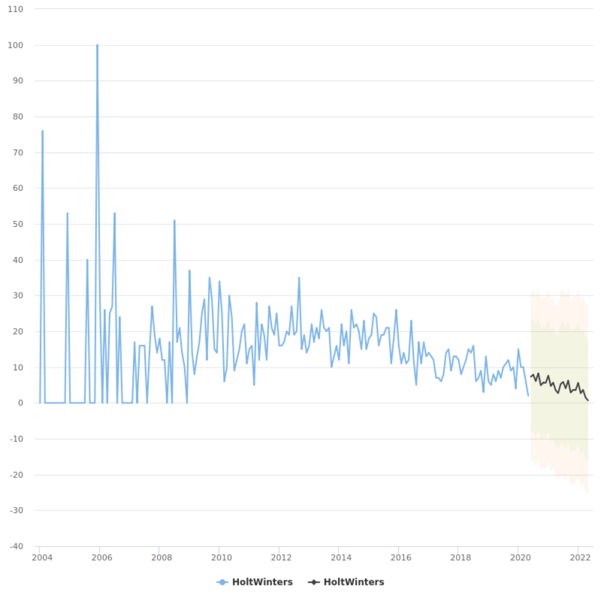


Figure 25: HoltWinters Forecast for University in Dubai, worldwide

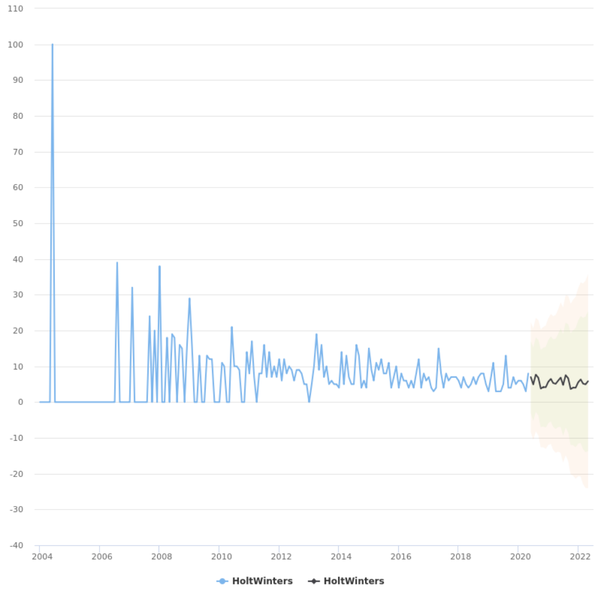


Figure 26: HoltWinters Forecast for University in Dubai, worldwide

6 Conclusion

This sections summarizes our findings for the project. We have consucted several analysis on 3 main datasets: KHDA census data, Google Reviews, Google Trends.

The aim of the first dataset was to analyse the correlation between several factors of the university data (Location, student numbers, ranking, other variables), and build a predictive model for linear regression to outline the student numbers for the future. We used Multi-linear, Quantile Regression models and Random Forest to build-up the approach. Unfortunately, due to quality of the dataset, it was not possible to construct a plug-and-play model, however we have achieved good results for RMSE and MAPE values.

The second analysis was conducted on Google Review dataset and included recommendation system (similar to our previous project, but applied on Google Reviews data via recosystem), as well as simple Sentiment Analysis of the provided reviews. The evaluation of our recommendation system was also done via RMSE and MAPE, both of which reached desired values. The Sentiment Analysis was performed for exploratory purposes.

The third analysis which we performed, was made on time-seried data of Google Trends from 2004 until date. The last set helped us to establish understanding where are we heading in this market considering the recent developments of COVID-19. The RMSE and MAPE was calculated for model.

Results - Summary on models

The below table records values for each algorithm used in the project, all data was divided into test and train sets prior to obtaining the final results.

Table 13: Final Project Results - Algorithms

Dataset	Algorithm	RMSE	MAPE
Census Data	Multi-Linear Reg	1.58	7.69
Census Data	Quantile Reg	1.51	1.84
Census Data	Random Forest	1.22	4.92
Google Reviews	Factorization	0.09	0.52
Google Reviews	Sentiment	N/A	N/A
Google Trends	Holt-Winters Forecast	14.8	17.3

Results - outcomes

Few very important outcomes which we have summarized for readers' convinience:

1. Collected data in the annual census is not representative to make any colclusion on the market activity. Aviability of data on higher education in Dubai is very limited and not well-maintained;
2. Major, Tuition Fees and how old the university is comprize significant factors for overall university success. The higher the values - the more successfull university is;
3. Universities receive mostly possitive reviews on Google Maps;
4. The main buzz-words for students leaving the reviews are: helpful, students, experience, quality, place, professors, people, money;
5. Online bachelors in Dubai, as a keyphrase, does not meet the minimum parameters as a Google Trend, both in the UAE and worldwide;
6. The overall trend for higher education in Dubai is decreasing since 2016, for both international and local markets;
7. COVID-19 might have insignificantly impacted on overall search volume in the last couple of months;
8. COVID-19 might have resulted in increase search volume for online master degrees in Dubai, both in UAE and worldwide;
9. The forecast for international and local students volume, who express interest to study in Dubai is negative;
10. The forecast for international and local students volume, who express interest to study online masters in Dubai is possitive.

Limitations

This project had significant limitations summarized as follows: limited access to detailed data for university census, limited demographics offered for google reviews data. The results of this analysis should be considered as “approximated”, however the build approach and models can be used to further enhance the development of the algorithms.

Further Research

It will be beneficial to expand the area of the research and improve several models presented. Specifically, obtain additional data which will present the opportunity to conduct the linear models with improved accuracy. A sentiment analysis can be improved by grouping words into groups and developing a machine learning algorithm to predict future reviews depending on the user’s demographics. This insights can be critial for universities to consider while developing their strategy for successful operations.

7 Appendix

7.1 List of Figurues and Tables

List of Figures

List of Tables

7.2 Session Info

R version 4.0.0 (2020-04-24)

Platform: x86_64-pc-linux-gnu (64-bit)

locale: LC_CTYPE=en_US.UTF-8, LC_NUMERIC=C, LC_TIME=en_US.UTF-8, LC_COLLATE=en_US.UTF-8, LC_MONETARY=en_US.UTF-8, LC_MESSAGES=C, LC_PAPER=en_US.UTF-8, LC_NAME=C, LC_ADDRESS=C, LC_TELEPHONE=C, LC_MEASUREMENT=en_US.UTF-8 and LC_IDENTIFICATION=C

attached base packages: stats, graphics, grDevices, utils, datasets, methods and base

other attached packages: rpart.plot(v.3.0.8), rpart(v.4.1-15), gdtools(v.0.2.2), tm(v.0.7-7), NLP(v.0.2-0), tidytext(v.0.2.4), syuzhet(v.1.0.4), wordcloud(v.2.6), RColorBrewer(v.1.1-2), gridExtra(v.2.3), recosystem(v.0.4.2), pander(v.0.6.3), skimr(v.2.1.1), svglite(v.1.2.3), quantreg(v.5.55), SparseM(v.1.78), randomForestExplainer(v.0.10.0), randomForest(v.4.6-14), CatEncoders(v.0.1.1), GGally(v.1.5.0), webshot(v.0.5.2), highcharter(v.0.7.0), forecast(v.8.12), reshape2(v.1.4.4), gtrendsR(v.1.4.6), dataPreparation(v.0.4.3), progress(v.1.2.2), Matrix(v.1.2-18), lubridate(v.1.7.8), knitr(v.1.28), kableExtra(v.1.1.0), data.table(v.1.12.8), caret(v.6.0-86), lattice(v.0.20-41), forcats(v.0.5.0), stringr(v.1.4.0), dplyr(v.0.8.5), purrr(v.0.3.4), tidyr(v.1.0.3), tibble(v.3.0.1), ggplot2(v.3.3.0), tidyverse(v.1.3.0), readr(v.1.3.1) and pacman(v.0.5.1)

loaded via a namespace (and not attached): readxl(v.1.3.1), backports(v.1.1.7), systemfonts(v.0.2.2), plyr(v.1.8.6), igraph(v.1.2.5), repr(v.1.1.0), splines(v.4.0.0), SnowballC(v.0.7.0), digest(v.0.6.25), foreach(v.1.5.0), htmltools(v.0.4.0), fansi(v.0.4.1), magrittr(v.1.5), recipes(v.0.1.12), modelr(v.0.1.7), gower(v.0.2.1), xts(v.0.12-0), tseries(v.0.10-47), prettyunits(v.1.1.1), colorspace(v.1.4-1), rvest(v.0.3.5), ggrepel(v.0.8.2), haven(v.2.2.0), xfun(v.0.13), callr(v.3.4.3), crayon(v.1.3.4), jsonlite(v.1.6.1), survival(v.3.1-12), zoo(v.1.8-8), iterators(v.1.0.12), glue(v.1.4.1), gtable(v.0.3.0), ipred(v.0.9-9), MatrixModels(v.0.4-1), quantmod(v.0.4.17), scales(v.1.1.1), DBI(v.1.1.0), Rcpp(v.1.0.4.6), viridisLite(v.0.3.0), stats4(v.4.0.0), lava(v.1.6.7), prodlim(v.2019.11.13), DT(v.0.13), htmlwidgets(v.1.5.1), httr(v.1.4.1), ellipsis(v.0.3.0), farver(v.2.0.3), pkgconfig(v.2.0.3), reshape(v.0.8.8), nnet(v.7.3-13), dbplyr(v.1.4.3), labeling(v.0.3), tidyselect(v.1.1.0), rlang(v.0.4.6), munsell(v.0.5.0), cellranger(v.1.1.0), tools(v.4.0.0), cli(v.2.0.2), generics(v.0.0.2), broom(v.0.5.6), evaluate(v.0.14), yaml(v.2.2.1), processx(v.3.4.2), ModelMetrics(v.1.2.2.2), fs(v.1.4.1), nlme(v.3.1-147), whisker(v.0.4), slam(v.0.1-47), xml2(v.1.3.2), tokenizers(v.0.2.1), compiler(v.4.0.0), rstudioapi(v.0.11), curl(v.4.3), reprex(v.0.3.0), stringi(v.1.4.6), highr(v.0.8), ps(v.1.3.3), urca(v.1.3-0), vctrs(v.0.3.0), pillar(v.1.4.4), lifecycle(v.0.2.0), lmtest(v.0.9-37), R6(v.2.4.1), janeaustenr(v.0.1.5), codetools(v.0.2-16), MASS(v.7.3-51.5), assertthat(v.0.2.1), withr(v.2.2.0), fracdiff(v.1.5-1), rlist(v.0.4.6.1), mgcv(v.1.8-31), parallel(v.4.0.0), hms(v.0.5.3), quadprog(v.1.5-8), grid(v.4.0.0), timeDate(v.3043.102), class(v.7.3-16), rmarkdown(v.2.1), TTR(v.0.23-6), pROC(v.1.16.2), base64enc(v.0.1-3) and tinytex(v.0.22)

Interested to collaborate on this project?

Get in touch via LinkedIn or **Check for updates on the project on GitHub**