Tanzania's Eco- tourism Development Research & Analysis by
Binh Minh An Nguyen & Concillia Hleziphi Mpofu

Overview

In this project we focused on eco-tourism development in Tanzania. As the tourism sector is on the road to recovery after severe effects from the COVID 19 pandemic, analyzing the possibility of a rebound while encouraging and optimizing its socio-economic development becomes more appealing and key in the current climate. The research and studies on tourism activities will help the local authorities and governments assess the viability of the sector and current economic incentives. Eco-tourism, is tourism to natural sites, wildlife, and reservations, it is becoming more and more appealing and critical economic sectors in many countries, including South East Asia, and African countries, Price (2017). Such a tourism segment not only directly contributes to the national economy itself but also expands the capacity and job markets, while also helping the locals to maintain and protect their ecosystem. Tourism is one of the significant contributors towards Gross Domestic Product for a number of developing nations around the world. According to the World Travel and Tourism Council, (2022), prior to the pandemic, Travel & Tourism (including its direct, indirect and induced impacts) accounted for 1 in 4 of all new jobs created across the world, 10.6% of all jobs (334 million), and 10.4% of global GDP (US$9.2 trillion).

Analytics Goals

- **Prediction goal**
  Predicting the revenue ("total_cost") based on the given tourism features and the tourists' demographic information
- **Inference goal**
  Examining the popular claim on tourism activities that:
  On average, Gen X-ers tend to spend more on tourism activities than any other age range does
- **Interpretation goal**
  Exploring and identifying which factors contribute most to Tanzania's ecotourism revenue

Primary Data ETL

The data was collected by the Tanzanian National Bureau of Statistics however we accessed the data from the Zindi Data Science community. The dataset contains 23 variables with 6476 distinct observations in total, which are splitted into a training set of 4810 rows and a test set of 1666 new observations.The dataset was collected to gain a better understanding of the status of the tourism sector and provide an instrument that will enable sector growth.

Prior to data analysis, we conducted primary data ingestion to rename the original variables in the Train.csv file, fix typo errors, and recode the values into a more readable format and insightful analysis. Since the majority of the predictors are categorical variables, we transformed the tourist's nationality into a numerical variable - traveling distance from the tourist's home

country to Tanzania by using the centroids among different countries. These centroids data is Google's Open Source data for community users.

Preliminary Analysis:

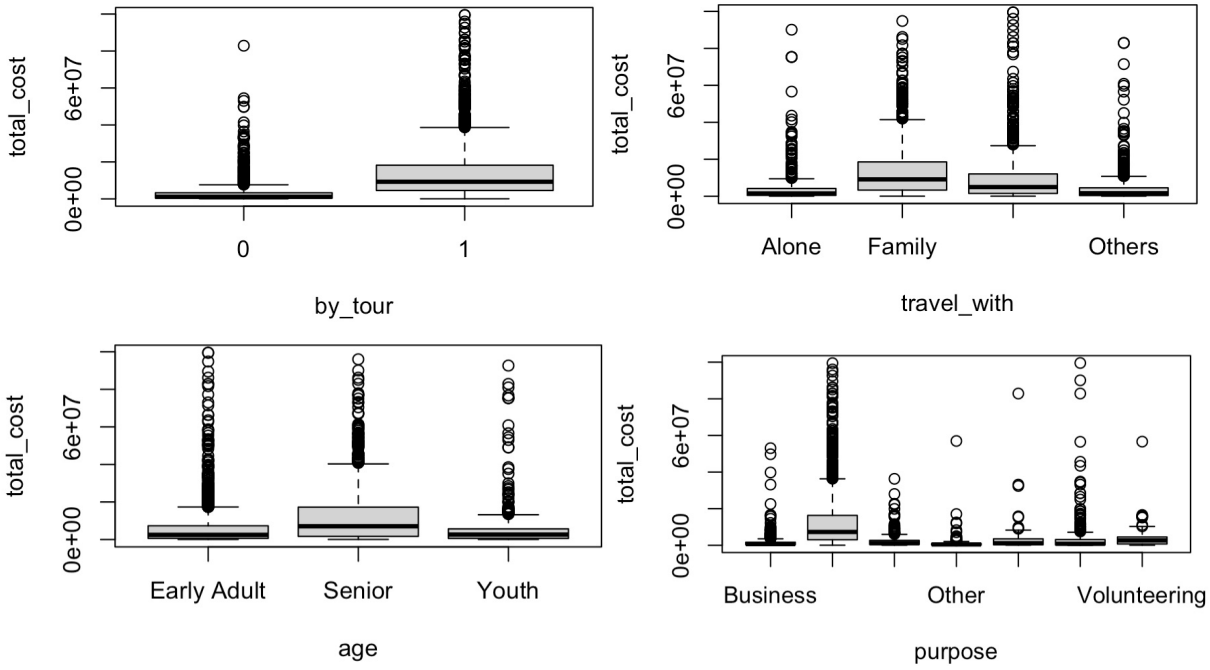
As the outcome variable is quantitative, total cost valued in Tanzanian Shillings, we went through examining all of 8 assumptions for the classical linear regression, and encountered 4 assumption violations, related to

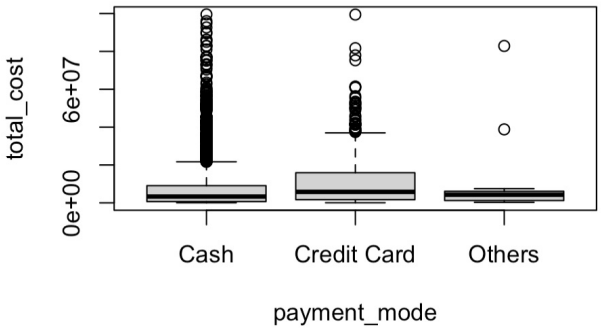
(1) A very weak to no linear relationship between outcome and predictors:

To inspect the potential relationships, we conducted an ANOVA test with a boxplot companion for examining if the outcome variable has a strong relationship with 15 categorical variables while using a correlation matrix for inspecting the correlation between the outcome and 5 quantitative variables.

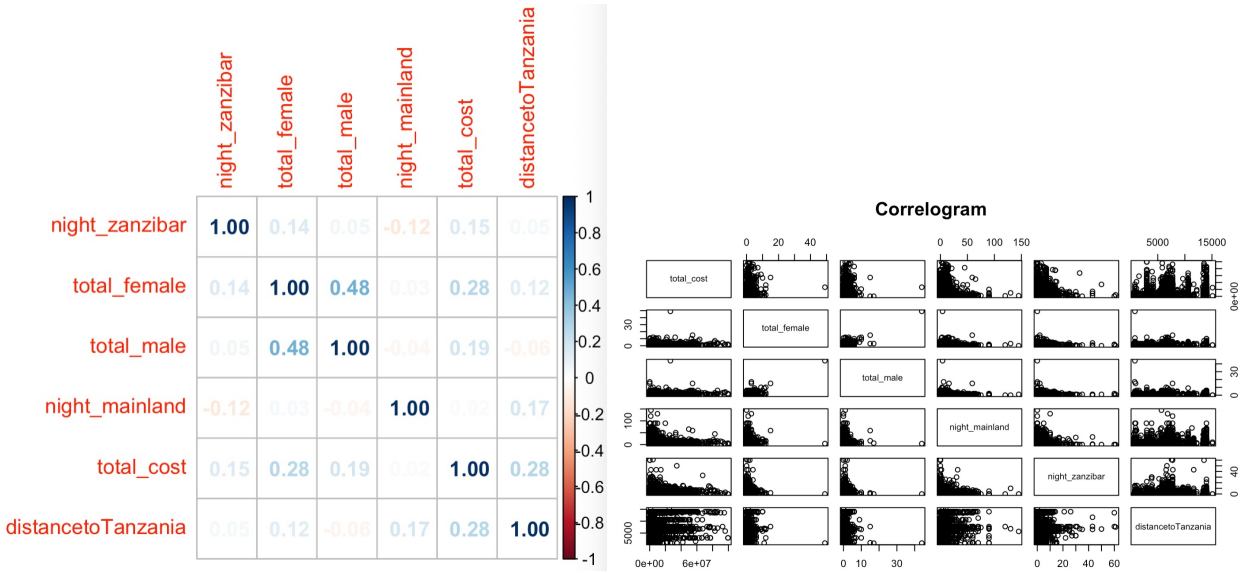
- **ANOVA Tests**

  We first conducted ANOVA test independently for each categorical variable against the outcome variable and visualized these tests with a box plot respectively. Both boxplot and single ANOVA test (based on the extremely small p-value) show that if standalone, all of these categorical variables are significant elements to predict the outcome total_cost. Nevertheless, as a rule of thumb, if a model includes more than 10 predictor variables, it's highly likely that the model will suffer from multicollinearity issues.





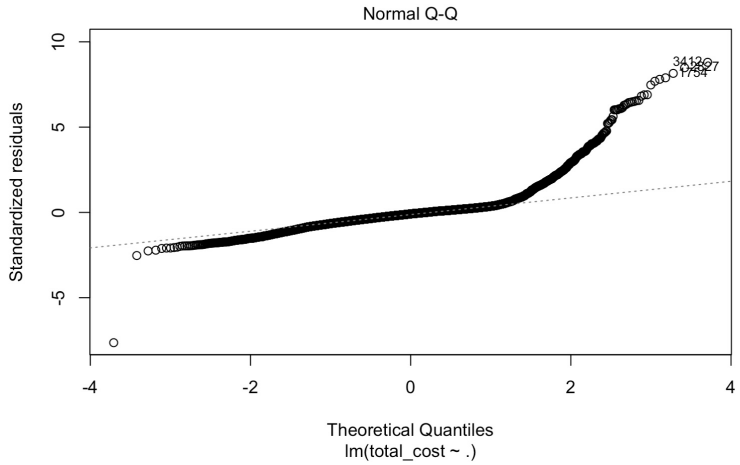
- **Correlations**
  The raw outcome variable total_cost does not hold any dramatic correlation with any of the quantitative variables. The highest correlation coefficient is only 0.28, belonging to total_cost against distance Tanzania. The correlogram plot also supports our observations, especially since there was an almost horizontal line representing the relationship between outcome and total_female and total_male.



Nevertheless, based on the boxplot, we can also see that several outliers are presented in our data. Besides, heteroskedasticity issue is also presented, ie. total_cost against night_mainland.

(2) Residuals term is not normally distributed

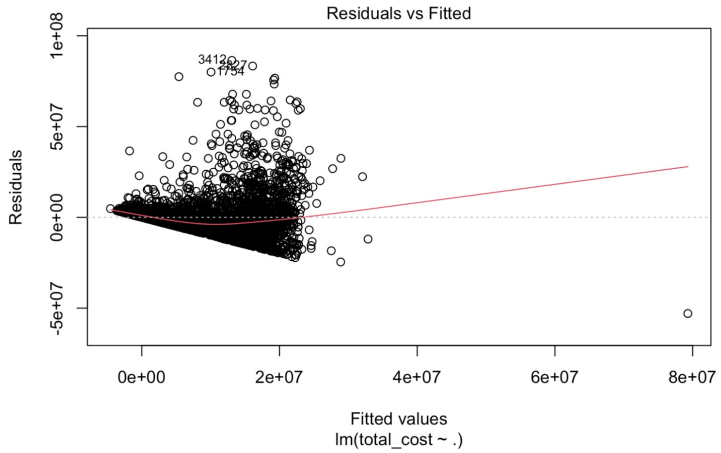
Based on the inspected relationship above, since the outcome still holds a strong relationship with the categorical variables, we decided to use the Ordinary Least Squares method to fit a linear regression line on our data using all 20 predictor variables. After that, we used plot() function to pull the residuals plots for analysis



The model returned that only 40% of the residuals term were overlapping the normal line in the normality plot. This proportion is far less than the given acceptance rate of 70% normality for medium-large size datasets with more than 50 degrees of freedom. Furthermore, the raw distribution of residuals term has both heavy tails departing from the normality line with several extreme outliers.

(3) Heteroskedasticity

As the residuals fan out when the fitted values increase, it indicates that the variance of the error term is not a constant. Nevertheless, from this plot, we also spotted several outliers



We further investigated heteroskedasticity by conducting 2 Test Statistics namely the Breusch-Pagan Test and Brown-Forsythe Test.

- Assumptions for Breusch-Pagan Test:
    - The sample size is large enough - here, the initial dataset with more than 4000 observations and 20 variables give us more than 50 degrees of freedom. Thus, this assumption is satisfied.
    - The residuals are independent and normally distributed - as mentioned above, the residuals term is not normally distributed in our case, which made the test being approximate and potentially misleading
    - The variance of disturbance term increases
- Assumption for Brown-Forsythe Test:

- The variance of disturbance term increases
- This is a non-parametric test, thus, it does not require normality for residuals. The test is based on the median and absolute differences between the residuals and their medians so is resistant to outliers.
- Requires a large enough sample size

The results are gathered below:

- Breusch-Pagan Test: BP = 377.91 and p-value < 2.2e-16
- Brown-Forsythe Test: BF = 7.374 and p-value = 0.0086

Given the small p-values, both Tests were significant and we have sufficient evidence to reject the NULL Hypothesis. Thus, we can conclude that heteroskedasticity was present.

(4) Severe multicollinearity

As mentioned above, when the model includes more than 10 variables, it's highly likely to involve multicollinearity. Thus, we firstly performed Chi-square Test to inspect the correlations among 15 categorical variables. Nevertheless, due to heteroskedasticity and outliers issues, our test statistic is just approximate. However, as most of the Chi-square Test returned with an extremely small p-value, the tests are significant and imply the red flags of severe collinearity issues in our data. The Chi-square Test showed that almost all of the categorical variables are strongly dependent on each other, especially for the by_tour (tour arrangement variable) and each of the package element variables. Such results are expected as the tour content should be decided in the first place, and it shall cover all costs related to the tour elements.

We further investigated the multicollinearity issue upon fitting an OLS Regression model by computing the Condition Index (CI) to assess the multicollinearity at the whole model level. The results for the Variance Inflation Index (VIF) for the variable level were as follows;

- **Condition Index**: it is given that if the CI <= 30, the collinearity is tolerable in our model. If 30 < CI <= 50, the multicollinearity is somewhat tolerable but requires consideration and further investigation. If the CI > 50, multicollinearity is severe in our model. For the OLS Regression model using the not-yet-transformed outcome variable, the highest CI is 26.46, which is lower than the first threshold of 30. Thus, at the model level, multicollinearity is tolerable.
- **Variance Inflation Factor**: it is given that if the VIF is associated with any variable in our models > 10, severe multicollinearity is involved. In our case, package_accomodation and by_tour have approximate VIF of 18 > 10, meaning the variance of these 2 variables is 18 times greater than it should be. Besides, there are several variables with VIF > 3, such as purpose, package_food, main_activity, and other package element variables.

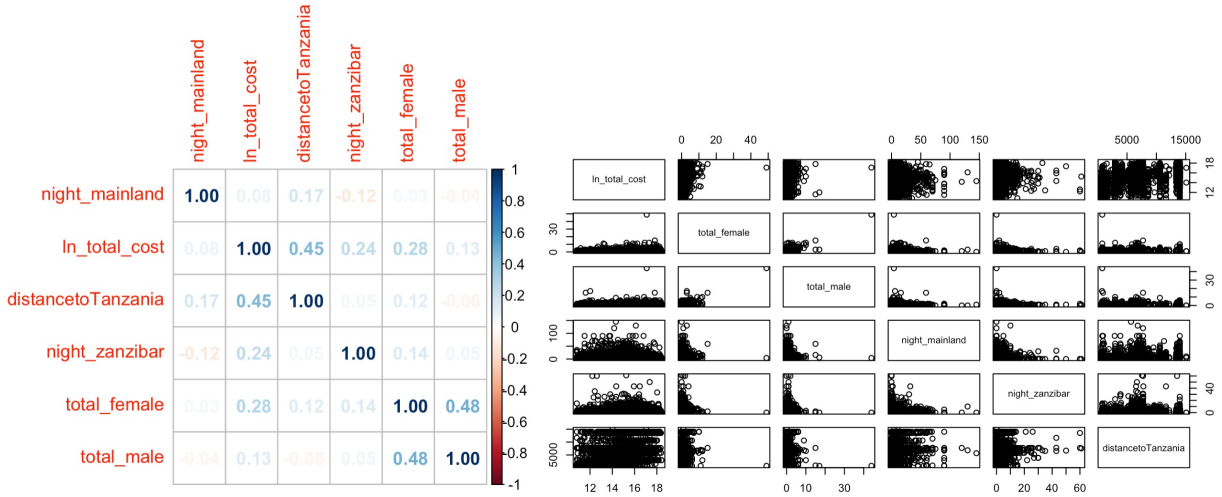
<u>Data Pre-processing (Further ETL)</u>

We performed 3-step data pre-processing to make remedies from these data issues. The further data ETL involves:

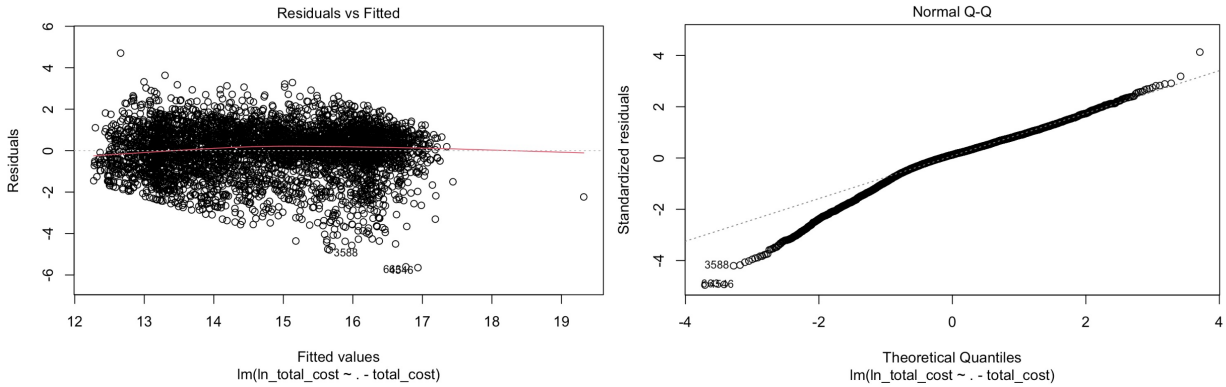
- Natural log transform outcome variable
- Remove outliers
- Overcome multicollinearity by variable selections

(1) Transform outcome variable

We used the natural log to transform the initial outcome variable total_cost into ln_total_cost = log(total_cost). The purpose of this transformation is to fix the heteroskedasticity issue and improve the relationship between outcome and predictor variables.



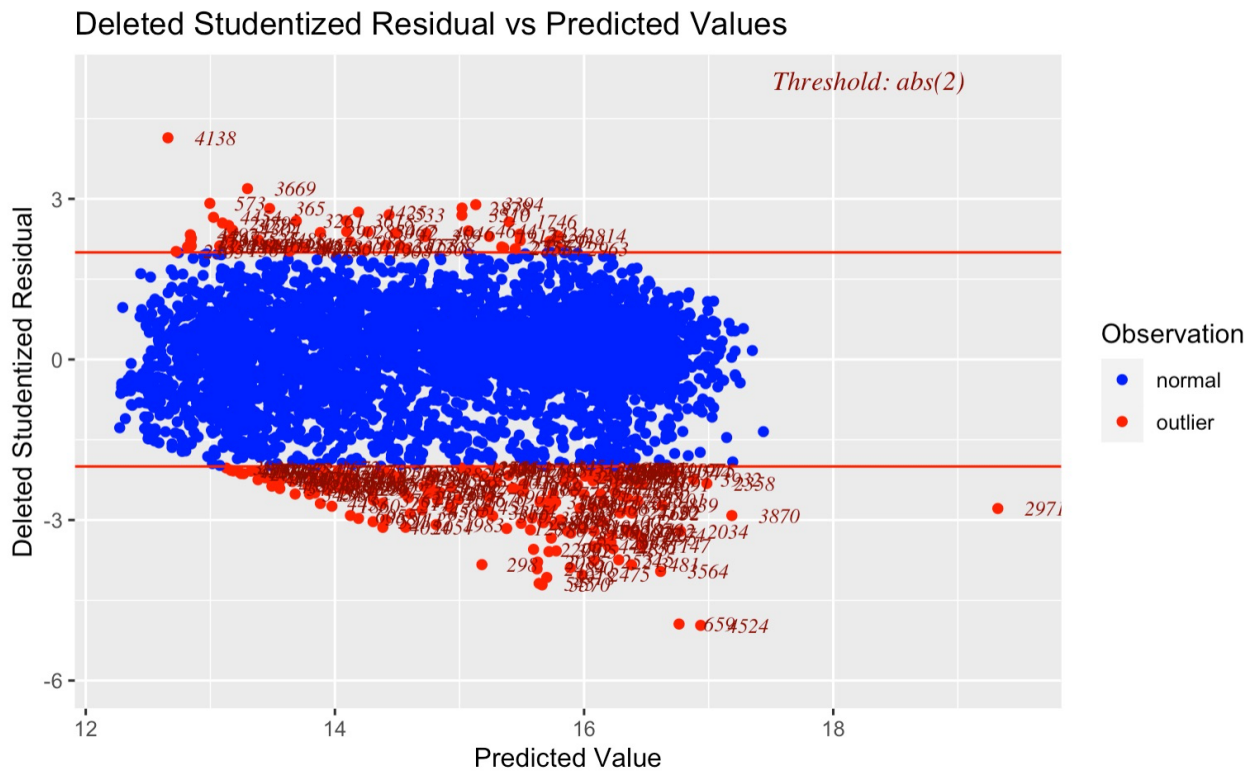
- Relationship between new outcome and predictors: the relationship between ln_total_cost and categorical variables are still significant.
- For the quantitative variables, while the relationship between ln_total_cost and total_male and total_female are improved, it's still difficult to interpret the relationship between outcome and night_mainland or night_zanzibar.



- The heteroskedasticity issue was fixed. We no longer observed any specific pattern on the residuals vs fitted value plot. Nevertheless, several outliers are still presented.

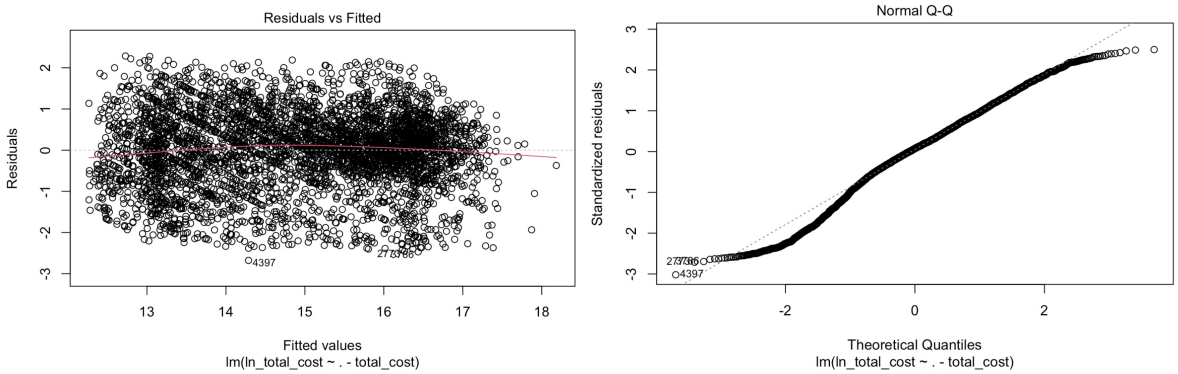
(2) Removing outliers for better data concentration

We applied the studentized residuals method with the threshold of absolute studentized residuals = 2. Any fitted values with the absolute studentized residuals value of more than 2 were excluded from our refined data. Once eliminating 101 unqualified observations, we extract the data again and store it in a variable named new_eco.



After that, we fit a regression model using the OLS method again on the new data and gather the result of residuals as below:

The full model after transforming the outcome variable is called Model 1 (mod1) with all 20 predictor variables.



After removing 101 outliers, the residuals plot became more neat and clear that there was no specific pattern in the residuals plot. In terms of normality, around 70% of the residuals term are now overlapping the normality line. For a medium-size dataset, this rate is acceptable to go with linear regression.

Nevertheless, while we acknowledged that there could be a structural break, we did not have enough time to investigate the issue, which variable created the break, and to make remedies for it. Thus, the team decided to go with the classical linear regression option.

(3) Overcome multicollinearity issue with stepwise variable selections

Upon fitting the full model mod1, we computed the CI and VIF indexes again:

- CI: the highest value is 26.7, indicating that at the model level, multicollinearity is tolerable.
- VIF: package_accomodation and by_tour still have the respective VIFs of 18.86 and 17.78, which is larger than the threshold of 10. Thus, multicollinearity is severe at variable level.

Since there was no restriction on the number of predictor variables, we assumed that some variables are allowed to drop from the full model. Since our data is pretty large with 20 variables, then we applied the stepwise variable selection loop to select the most significant variables only. The mechanism of the stepwise process is similar to using an ANOVA breakdown table to compare the significance of each variable based on the magnitude of their p-values. The stepwise method takes the full model (mod1) as the upper limit, and the NULL model (mod2) as the lower limit, then runs both step-backwards and step-forward processes. Once at a time, the most insignificant variable will be removed from the model. Then each will be added back to the model again to check if their significance changes. The process will stop once there is no more insignificant variable to be removed.

At the end of this process, we got a new specification of 15 variables including both quantitative and categorical variables. Nevertheless, we realized that first_trip_tz is only significant at the magnitude of p-value < 0.1. Thus, we remove this variable from our new specification then fit the OLS Regression again with 14 variables (mod2). Compared to mod1, mod2 specification excludes info_source, first_trip_tz, and 4 package element variables. After that, we computed the CI and VIF index again:

- CI: The highest value of the condition index is now only 14.92, which is much lower than that of the full model mod1 and indicates that the multicollinearity is tolerable at the model level.
- VIF: None of the VIFs are higher than 10, suggesting that at the variable level, multicollinearity is tolerable in general. However, since there are several variables with associated VIF less than 10 but more than 3, we are also interested in more robust regression to deal with such high variance inflation factors, such as Ridge and Principle of Component (PCR).

We will further evaluate the goodness of fit of each model in the Test Statistic section below.

<u>Test Statistic</u>

1. **Test Statistic for Age**
   H0:$\beta_{age}=0$

   Ha:$\beta_{age}\neq 0$

   T Statistic = 3.55119299

   P value = 3.874192e-04

Conclusion:
Given the test statistic = 3.55119299 and an associated p value = 3.874192e-04, there is overwhelming statistical evidence to indicate that people at different age ranges will spend

differently on traveling. Thus, we can reject the NULL Hypothesis.
Nevertheless, our test cannot fully support the statement that average Gen X-ers tend to spend more on tourism activities than any other age range.

2. **Model's Goodness of Fit Test**

We compare the goodness of fit between 2 models based on 3 criteria: (1) Adjusted R-squared, (2) AIC and BIC statistical indexes, (3) Extra Sum of Square F-test

models <chr>	AIC <dbl>	BIC <dbl>	r.squared <dbl>	adj.r.squared <dbl>
mod1	12103.42	12360.12	0.6767775	0.6740395
mod2	12098.79	12284.90	0.6755345	0.6735864

2 rows | 1–10 of 13 columns

a. Adjusted R-squared

The adjusted R-squared of the full model (mod1) is just 0.4% higher than that of the reduced model (mod2) with 14 variables. Generally, we will select the model with a higher value of Adjusted R-squared - mod1. Nevertheless, given the fact that the full model (mod1) has severe multicollinearity issues, it is tolerable in model 2. Thus, we select model 2 as the optimal model.

b. AIC and BIC statistical indexes

Model 2 has lower values of both AIC (12098.79) and BIC (12284.90) compared to the full model, in which AIC = 12103.42 and BIC = 12360.12. Thus, we select model 2 with lower AIC/BIC indexes.

c. Goodness of Fit F-test

To justify if the complexity of the full model with extra 6 variables is worths the model's explanatory power, we have ANOVA breakdown table for model 1 in addition to the ANOVA direct test to compare these 2 models.

H0: coefficients of these 6 variables are 0
Ha: at least one of the coefficients is different from 0.

Alternative:
Full mode = model 1 with 20 variables
Reduced model = model 2 with 14 variables, excluding first_trip_tz, info_source, package_accommodation, package_sightseeing, package_insurance, package_guided_tour

- Model 1 - ANOVA breakdown table

	Source	Df	SS	MS	F	P
	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
11	payment_mode	2	1.149368e+02	57.4684067	68.2615021	6.259822e-30
12	age	2	1.053499e+02	52.6749689	62.5678126	1.582085e-27
13	by_tour	1	5.429252e+01	54.2925152	64.4891490	1.229826e-15
14	distancetoTanzania	1	2.804829e+02	280.4828666	333.1601293	7.420599e-72
15	info_source	6	7.967440e+00	1.3279067	1.5772998	1.494092e-01
16	first_trip_tz	1	2.391148e+00	2.3911481	2.8402277	9.200073e-02
17	package_accomodation	1	1.614910e+00	1.6149103	1.9182054	1.661235e-01
18	package_sightseeing	1	1.328740e+00	1.3287395	1.5782890	2.090727e-01
19	package_guided_tour	1	1.006123e+00	1.0061227	1.1950818	2.743656e-01
20	package_insurance	1	2.152897e-01	0.2152897	0.2557231	6.130987e-01

We first resequenced the variables in model 1 to follow those in model 2 then added these 6 variables (info_course, first_trip_tz, and 4 package element variables) at the end of the list. Given the associated p-value > 0.05, all of these variables are not as significant as the other 14 variables.

- ANOVA test to compare 2 models

```
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1   4497 3791.2
2   4486 3776.7 11    14.524 1.5683 0.1011
```

Conclusion: The test statistic F = 1.5683 and a p-value of 0.1011, our test is not significant. Thus, there is insufficient evidence for us to reject the NULL Hypothesis. Therefore, we conclude that adding 6 variables does not contribute to the model statistical explanatory power. Following the Principle of Parsimony, we, thereby, select model 2 as the more superior model.

Model Building Process

1. **Variable selection**

As mentioned above, we will use the mod2 specification for further model building and validations. The new specification contains 14 significant variables (both quantitative and categorical variables) while excluding info_source, first_trip_tz, and 4 package element variables.

2. **State the Initial model you considered**

We will use Ordinary Least Square Regression using the new specification, Ridge Regression, and Principle of Component Regression.

For each regression method, we will build and validate the model by using the 10-fold cross-validation method. This method allows us to train and test our model at the same time. The method will randomly split the data into 10 folds or 10 subsets. Every time, 9 subsets will be taken to build the model, then the model will be tested against the left 1 subset. The process is completed once all data points are used for building and testing the model.

3. Model Building and Validation

a. Ordinary Least Squares Regression (OLS)

```
                                      Estimate Std. Error t value Pr(>|t|)
(Intercept)                           1.286e+01  5.512e-02 233.232  < 2e-16 ***
travel_withFamily                     4.443e-01  4.414e-02  10.067  < 2e-16 ***
`travel_withFriends/Relatives`        3.327e-01  4.848e-02   6.864 7.60e-12 ***
travel_withOthers                    -1.031e-01  3.960e-02  -2.603 0.009263 **
total_female                          1.238e-01  1.503e-02   8.238 2.28e-16 ***
total_male                            5.622e-02  1.688e-02   3.330 0.000874 ***
`purposeLeisure and Holidays`         5.570e-01  5.521e-02  10.090  < 2e-16 ***
`purposeMeetings and Conference`      1.423e-01  7.815e-02   1.820 0.068805 .
purposeOther                         -8.743e-02  9.860e-02  -0.887 0.375311
`purposeScientific and Academic`     -1.318e-01  1.177e-01  -1.120 0.262765
`purposeVisiting Friends and Relatives` -1.113e-01  6.012e-02  -1.851 0.064279 .
purposeVolunteering                   3.053e-01  1.027e-01   2.972 0.002972 **
`main_activityBusiness tour`          1.753e-01  5.844e-02   3.000 0.002711 **
`main_activityCultural tourism`      -3.546e-01  6.025e-02  -5.886 4.25e-09 ***
`main_activityHunting tourism`       -3.813e-01  6.243e-02  -6.106 1.10e-09 ***
`main_activityMountain climbing`     -1.402e-01  8.774e-02  -1.598 0.110151
`main_activityWildlife tourism`       8.171e-02  4.011e-02   2.037 0.041715 *
package_transport_intYes              4.252e-01  4.330e-02   9.821  < 2e-16 ***
package_foodYes                       2.148e-01  6.926e-02   3.101 0.001939 **
package_transport_tzYes               1.330e-01  5.537e-02   2.402 0.016340 *
night_mainland                        2.206e-02  1.472e-03  14.987  < 2e-16 ***
night_zanzibar                        4.242e-02  3.913e-03  10.840  < 2e-16 ***
`payment_modeCredit Card`             3.576e-01  4.192e-02   8.529  < 2e-16 ***
payment_modeOthers                   -5.930e-02  2.381e-01  -0.249 0.803360
ageSenior                             2.202e-01  3.185e-02   6.912 5.45e-12 ***
ageYouth                             -2.003e-01  4.590e-02  -4.363 1.31e-05 ***
by_tour1                              5.403e-01  8.095e-02   6.675 2.78e-11 ***
distancetoTanzania                    7.386e-05  4.050e-06  18.240  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- Model interpretation:
  - The coefficient of Age is significant. On average, keeping every other factor constant, Seniors tend to spend more on tourism than the Early Adults by 22%, while the Youth tends to spend less by 20%.
    It is interesting that there was a gap age group in the original dataset, by which none of the observations belong to the Late Adult age group (45-65 years old). While this age group is part of Gen X-ers, we don’t have enough data for our analysis to verify the claim on tourism activities. Nevertheless, the results in our coefficient are understandable as Youth include children, teenagers, and students who do not have a stable income, and seek less comfort while traveling.
  - On average, keeping everything else constant, a person will spend more than solo traveling by 44.4% if it is a family trip, and by 33.3% if it is a friends reunion, and spend less by 10.31% if the person has to travel with other non-relative people.
  - On average, keeping everything else constant, if there is one more male traveling in the group, the total_cost will increase by 5.6% while it increases by 12.3% if there is one more female traveling. Such a result is explainable as ladies tend to spend more on comfort than men

intercept	RMSE	Rsquared
<lgl>	<dbl>	<dbl>
TRUE	0.9221523	0.6710893

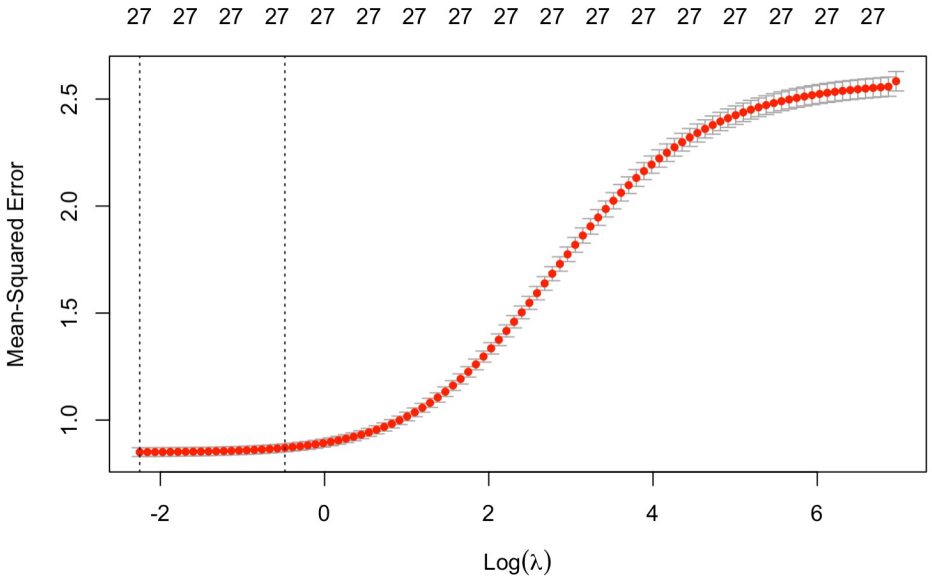
- Given the R-squared of 67.10%, the model does a pretty good job at explaining the variability of the outcome variable ln_total_cost
- The Best RMSE = 0.92215 - we used this cross-validation RMSE in the model comparison and selection below.

b. Ridge Regression

```
(Intercept)                          1.293359e+01
travel_withFamily                    4.169759e-01
travel_withFriends/Relatives         2.997511e-01
travel_withOthers                   -1.183465e-01
total_female                         1.279602e-01
total_male                           5.823625e-02
purposeLeisure and Holidays          5.154285e-01
purposeMeetings and Conference       9.763970e-02
purposeOther                        -1.473976e-01
purposeScientific and Academic      -1.509696e-01
purposeVisiting Friends and Relatives -1.471125e-01
purposeVolunteering                  2.590192e-01
main_activityBusiness tour           1.797253e-01
main_activityCultural tourism       -3.295865e-01
main_activityHunting tourism        -3.693579e-01
main_activityMountain climbing      -1.394182e-01
main_activityWildlife tourism        8.186436e-02
package_transport_intYes             4.188922e-01
package_foodYes                      2.627639e-01
package_transport_tzYes              1.816043e-01
night_mainland                       2.067598e-02
night_zanzibar                       4.105587e-02
payment_modeCredit Card              3.416909e-01
payment_modeOthers                  -4.473170e-02
ageSenior                            2.177412e-01
ageYouth                            -1.849174e-01
by_tour1                             4.564916e-01
distancetoTanzania                   7.134726e-05
```

- Model Interpretation:
  We used the Ridge Regression to minimize the effects of multicollinearity in our model using a tuning parameter lambda to determine how much to shrink the coefficients. The larger lambda, the more shrinkage is applied; however, the coefficients will never be shrunk to exactly 0. Here, the effects of predictors in the Ridge model with the Best Lambda are quite similar to those of the OLS Regression model. Since the best shrinkage lambda is small, only 0.105.
  - On average, keeping everything else constant, Seniors tend to spend more on tourism than the Early Adult by 21.77%, while the Youth tends to spend less by 18.49%.

- On average, keeping everything else constant, when the person travels with family, he/she will be willing to spend more than when traveling alone by 41.70%, and by 30% when traveling with friends or relatives, while he/she will spend less by 11.83% if the person has to travel with non-relative people.
- On average, keeping everything else constant, when there is one more female in the travelling group, the total cost will increase by 12.80%, while it is 5.8% when there is one more male.

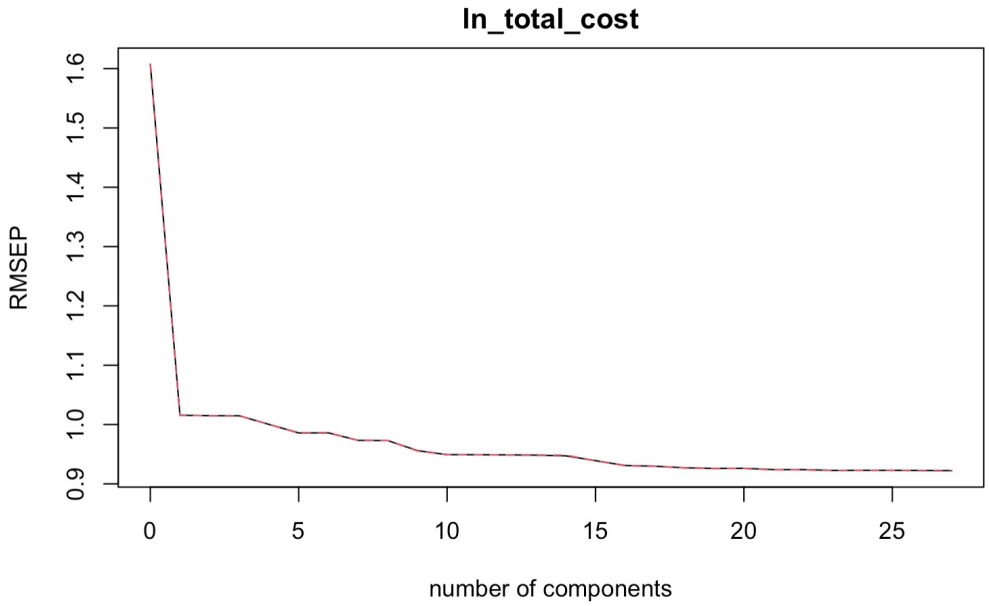



```
Best Ridge Lambda Best Ridge MSE Best Ridge RMSE
        0.1053521      0.8500239       0.9219674
```

- For Ridge Regression, we can compute the cross-validation RMSE of 0.9219, while we encounter a computation constraint in calculating the R-squared values since we use L2 Norm regularizations. Nevertheless, we will use the computed cross-validation RMSE for model comparison later.

c. Principle of Component Regression (PCR)

Different from Ridge, the Principle of Component takes components instead of variables to build a regression model. Thus, all collinearity issues are removed totally. In our case, there are 27 models with 27 components returned. The result is shown below:




```
Data:   X dimension: 4525 27
        Y dimension: 4525 1
Fit method: svdpc
Number of components considered: 27

VALIDATION: RMSEP
Cross-validated using 10 random segments.
       (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps
CV           1.607    1.016    1.015    1.015        1   0.9859
adjCV        1.607    1.016    1.015    1.015        1   0.9858
       6 comps  7 comps  8 comps  9 comps  10 comps  11 comps
CV       0.986   0.9734    0.973   0.9561    0.9492    0.9491
adjCV    0.986   0.9733    0.973   0.9554    0.9491    0.9488
       12 comps  13 comps  14 comps  15 comps  16 comps  17 comps
CV       0.9488    0.9484    0.9472    0.9391    0.9309    0.9297
adjCV    0.9487    0.9483    0.9485    0.9388    0.9305    0.9295
       18 comps  19 comps  20 comps  21 comps  22 comps  23 comps
CV       0.9270    0.9260    0.9262    0.9240    0.9240    0.9227
adjCV    0.9268    0.9257    0.9259    0.9236    0.9237    0.9223
       24 comps  25 comps  26 comps  27 comps
CV       0.9228    0.9228    0.9226    0.9223
adjCV    0.9225    0.9225    0.9222    0.9220

TRAINING: % variance explained
              1 comps  2 comps  3 comps  4 comps  5 comps  6 comps
X               18.54    25.83    32.57    38.72    44.34    49.47
ln_total_cost   60.05    60.13    60.16    61.34    62.47    62.47
              7 comps  8 comps  9 comps  10 comps  11 comps
X               54.04    58.24    62.18     66.00     69.71
ln_total_cost   63.45    63.51    64.87     65.27     65.35
              12 comps  13 comps  14 comps  15 comps  16 comps
X                73.29     76.49     79.41     82.30     84.98
ln_total_cost    65.35     65.41     65.49     66.14     66.79
              17 comps  18 comps  19 comps  20 comps  21 comps
X                87.44     89.76     91.89     93.74     95.24
ln_total_cost    66.85     67.07     67.16     67.16     67.34
              22 comps  23 comps  24 comps  25 comps  26 comps
X                96.66     97.63     98.50     99.18     99.71
ln_total_cost    67.35     67.47     67.47     67.49     67.52
              27 comps
X               100.00
ln_total_cost    67.55
```

- Looking at the SCREE plot, we can see a very clear elbow at component = 1, a good elbow at component = 5, and some faint elbows at component = 10 and 16. Nevertheless, the RMSE line keeps going down overall, and the lowest point is at component = 27. Nevertheless, we will further examine which model with the number of components to be our optimal choice based on 3 criteria below:

  Cross-validation RMSE: when component = 27, the model returns the lowest RMSE result of 0.9223. Given this result, the model with 27 components will be the best model for predictive accuracy purposes.

  The proportion of explained variance of predictor variables, in general, will increase as we add more components to the model. However, 70% is the general acceptance rate for any PCR model. Here, from the point where component = 12 onward, the models pass this acceptance threshold. 100% of the variance of predictors is explained by the model with 27 components.

The proportion of explained variance in the outcome variable is equivalent to the R-squared value. As component = 27, the model can best explain the variability of the outcome by 67.55%.

- In general, the PCR model is good for prediction purposes, while it uses the component approach instead of variables, the regression method does not support interpretation as much as other regression methods.

- Model Interpretation:

	12 comps	27 comps
travel_withFamily	0.1806799812	0.207521641
travel_withFriends/Relatives	0.0430746882	0.128908802
travel_withOthers	-0.1502188457	-0.043368581
total_female	0.1755049041	0.132354372
total_male	0.0866252910	0.051291280
purposeLeisure and Holidays	0.1989488770	0.274058011
purposeMeetings and Conference	-0.0221667516	0.035121264
purposeOther	-0.1052575684	-0.014330080
purposeScientific and Academic	0.0055627599	-0.017589351
purposeVisiting Friends and Relatives	-0.1078409533	-0.037871880
purposeVolunteering	0.0700502772	0.049060677
main_activityBusiness tour	0.1071539710	0.049725319
main_activityCultural tourism	-0.0415763255	-0.093107919
main_activityHunting tourism	-0.1177204497	-0.111583757
main_activityMountain climbing	-0.0206762933	-0.030414898
main_activityWildlife tourism	0.0447095529	0.040817091
package_transport_intYes	0.1280911213	0.195026854
package_foodYes	0.1641404196	0.106249713
package_transport_tzYes	0.1659148065	0.064933408
night_mainland	0.1286009851	0.225430065
night_zanzibar	0.0758172244	0.171694477
payment_modeCredit Card	0.1372747584	0.119802971
payment_modeOthers	0.0071717680	-0.003408965
ageSenior	0.1420584243	0.105419844
ageYouth	0.0002319119	-0.066766341
by_tour1	0.1800798803	0.269487811
distancetoTanzania	0.2420639306	0.314118839

Firstly, we recognize that the PCR model does not return any y-intercept, thus as expected, the regression coefficients change much compared to the other 2 models. Nevertheless, we still observe similar coefficient effects as other models. In particular:

- On average, keeping everything else constant, Seniors tend to spend more on tourism than the Early Adult by 10.54%, while the Youth tends to spend less by 6.68%.
- On average, keeping everything else constant, when the person travels with family, he/she will be willing to spend more than when traveling alone by 20.75%,

and by 12.89% when traveling with friends or relatives, while he/she will spend less by 4.36% if the person has to travel with non-relative people.

- On average, keeping everything else constant, when there is one more female in the traveling group, the total cost will increase by 13.24%, while it is 5.13% when there is one more male.

4. Model Selections for Analytics Goals

Rounding back to our 3 analytics goals:

- Prediction Accuracy: To predict the total_cost spent by a person given their demographic information and tour elements.
  The general rules for model selection here is to select the model that returns the lowest value of cross-validation RMSE. As shown below, Ridge Regression with the best lambda of 0.105 returns the lowest cross-validation RMSE out of the 3 models. Thus Ridge Regression with the Best lambda is the most optimal choice for prediction accuracy purposes.

	RMSE
Best Ridge	**0.9219**
Best OLS	0.9222
Best PCR	0.9223

- We can choose either the Ordinary Least Squares Regression model or any Ridge Regression model with an acceptable small lambda for interpretation and inference purposes. Here, an example for the interpretation purpose, we can conclude that the Tanzanian Government, local authorities, tourism development bodies can target on those tourists with the following characteristics:
  - Age range: adult to senior - since these groups of tourist are who normally has stable incomes, and seeking comfort while traveling. Thus, they are willing to spend more on the quality of the trip than just having budget experience
  - Nature of the trip (whom to travel with): recommending tours and target on group of family, relatives, or friends
  - Purpose of the ecotour: for Leisure and Holidays, and Volunteer - those that allow high degree of relaxation activities
  - Payment method: to focus on people who choose to pay by either Credit Card or cash, since these 2 payment methods can be proceeded anytime out of plan, while the service provider can earn the revenue immediately.
- Inference: As we use the Ordinary Least Square Regression model 2 to conduct the test statistic for the significance of Age toward predicting the outcome variable ln_total_cost earlier. Based on the test result, we can only conclude that Age is a significant predictor in our model, while we cannot verify the claim regarding the tourism activities by Gen X-ers.

Discussion

- As we use random split data in our analysis, we expect to have some differences in modeling results as if anyone takes our Rcode and rerun the modeling and analysis. In fact, we encounter this issue among our team members. For instance, after removing outliers from the dataset, Connie ran the model with only 1 dummy variable of Senior age group, while An's original codes return 2 dummy variables of Youth and Seniors, upon the reference level = Early Adult. Therefore, we would like to raise awareness when others replicate our research.
- The use of the initial country variable: Initially, we planned to group the tourist's home country column into 2 categories of developed and developing countries and examine if the tourists from developed countries tend to have luxury tourism. However, since the original dataset has already involved 15 categorical variables and we already had another inference goal of testing the claim on Gen X-ers, having another inference goal on different categorical variables would make our analysis more dull. Thus, we decided to go with the suggestion to convert this column into a numerical variable - the traveling distance to Tanzania. As a result, we have a new analytics goal of interpretation to identify which are the significant factors for ecotourism.
- As mentioned earlier, we can hardly interpret if there was a linear relationship between the predictor variable and some quantitative variables, such as the Distance to Tanzania, number of female or male traveling, even after making data transformation. In addition, there is a break in the lower tail of the residuals distribution. Thus, we suspected if other regression models can represent our data better, perhaps by including interaction terms or having structure breaks addressed in the final model. If more time is given, our team will definitely investigate and resolve such suspicions.
- Besides, as part of the inference goal - to examine if Gen X-ers spend more on traveling than any other age group, we are also interested in gathering more information on the spending patterns of the Late Adult age group.
- In terms of computation limitation, we failed to calculate the adjusted R-squared for Ridge Regression model, the interval for the final models selected, and conduct test statistics on the best model. In this research, we have calculated these elements using the model 2 - which was built on the whole dataset without model testing or validation. As these final models were built with R machine learning packages: {caret} package for the cross-validation on OLS model, {boot} package for Ridge Regression cross-validation, and {pls} package for PCR model. As the result and in addition to the time constraint, we ran out of resources to resolve these issues.
- Last but not least, because of the shortage of time and resources, we were not able to perform a full life cycle of data analytics, following the CRISP-DM model (cross-industry standard process for data mining). Compared to the standard process, our research and analysis is currently stopped at half-way of step 5 - evaluate models without recycling.

[Reference: https://commons.wikimedia.org/wiki/File:CRISP-DM_Process_Diagram.png]

Group Work

We held a stand meeting every Tuesday at 12.30PM to discuss the project's next step and assign appropriate tasks to each member. Due to the short time constraint to complete the group project, the tasks are splitted based on each team member's strengths. Both team members agreed to freely take the lead on the tasks that we were comfortable with.

- Connie with her effortless but powerful writing skills will take the lead on creating the presentation and supervise any writing works.
- An took lead on directing the project, coding and designing project content.

We used Google shared drive as our single point of collaboration, while staying open to any communication method. Throughout the project, we did not face any issue in terms of communication and task completion.

References

1. Price, R.A. (2017). The contribution of wildlife to the economies of Sub Saharan Africa. K4D Helpdesk Report. Brighton, UK: Institute of Development Studies, Accesed from https://assets.publishing.service.gov.uk/media/59ad5313ed915d78233d6474/145-Contribution-of-wildlife-to-SSA-economies.pdf
2. Tanzania Tourism Prediction by Pycon Tanzania Community accessed from https://zindi.africa/competitions/tanzania-tourism-prediction/data?fbclid=IwAR3UD66uc8KiKj8p4_4rSfrUW77z8dveJTrsVhkDpiOsZtUJgRYtk8s1B2s
3. World Travel Tourism, 2022, Economic Impact Reports, Accessed from https://wttc.org/Research/Economic-Impact
4. Zhen-yu Mei et al. (2019) Research on a forecasting model of tourism traffic volume in theme parks in China, Transportation Safety and Environment, 2019, Vol. 1, No. 2