# Tanzania’s Eco- tourism Development Research & Analysis

Binh Minh An Nguyen  & Concillia Hleziphi Mpofu

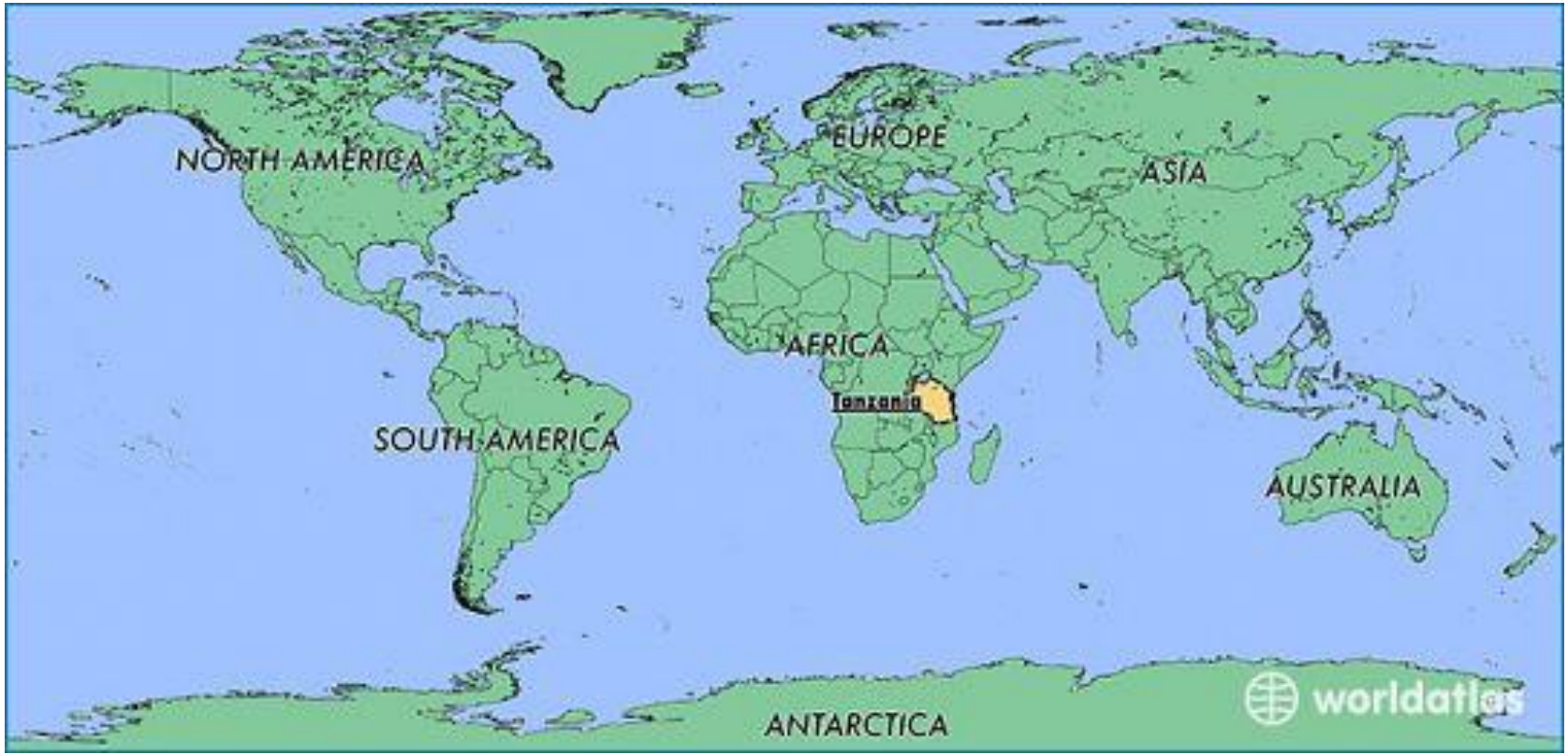
# RESEARCH OVERVIEW

- Data Source: the Zindi African Data Scientist Community.
- Surveillance data from seven points, from Tanzania where the eco-tourism activities were conducted.

**Motivation**

- Tourism sector is one of the sectors severely affected by the covid 19 pandemic and the sector is in need of prediction models in their road to discovery.
- Tourism in a major contributor towards GDP in most African countries

*According to Hussaine Jummane 2021, tourism is one of the largest sectors in Tanzania and it contributes 17.2% towards the national GDP and 25% of all foreign exchange revenues. The sector also provides employment for more than 600,000 people in the country.*


NORTH AMERICA
EUROPE
ASIA
AFRICA
Tanzania
SOUTH AMERICA
AUSTRALIA
ANTARCTICA
worldatlas

# RESEARCH 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

# DATASET OVERVIEW

- Data from Tanzania National Bureau of Statistics
- 23 variables with 4810 observations
- Outcome: total_cost in TZS Currency
- **Data Types**
  *Quantitative*: total_cost, number of people, number of nights
  *Categorical/Binary*: details about tours, tourist demographic information
  *Additional predictor:* Distance travelled by tourists using the centroids from google maps
- **Purpose**: To gain a better understanding of the status of the tourism sector and provide an instrument that will enable sector growth.

# PRELIMINARY ANALYSIS

## Violation of Assumptions

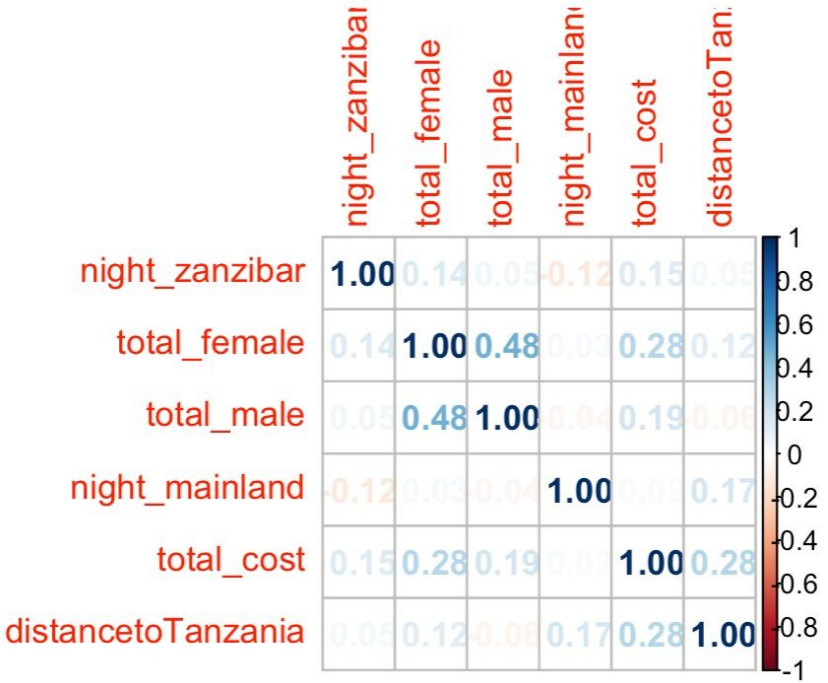
- Faint linear relationship between outcome and predictors
- Residuals term is not normally distributed
- Heteroskedasticity
- Multicollinearity

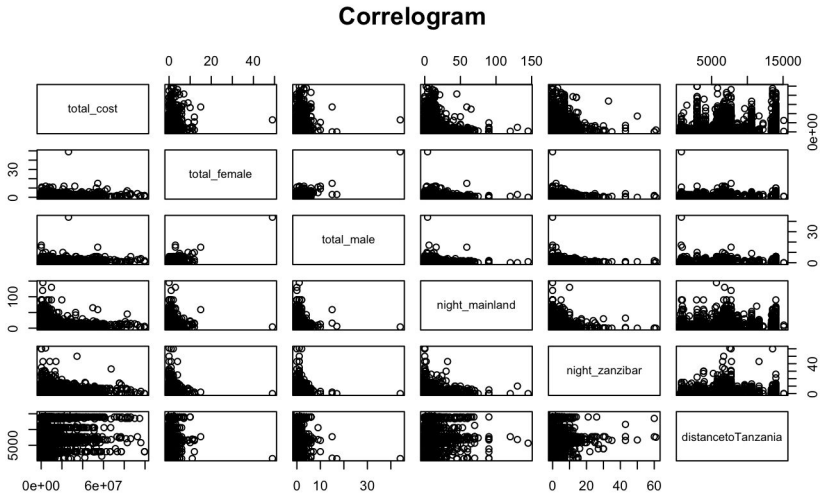
## Others

- Outliers

# PRELIMINARY ANALYSIS

## Relationship between Outcome and Predictors



# PRELIMINARY ANALYSIS

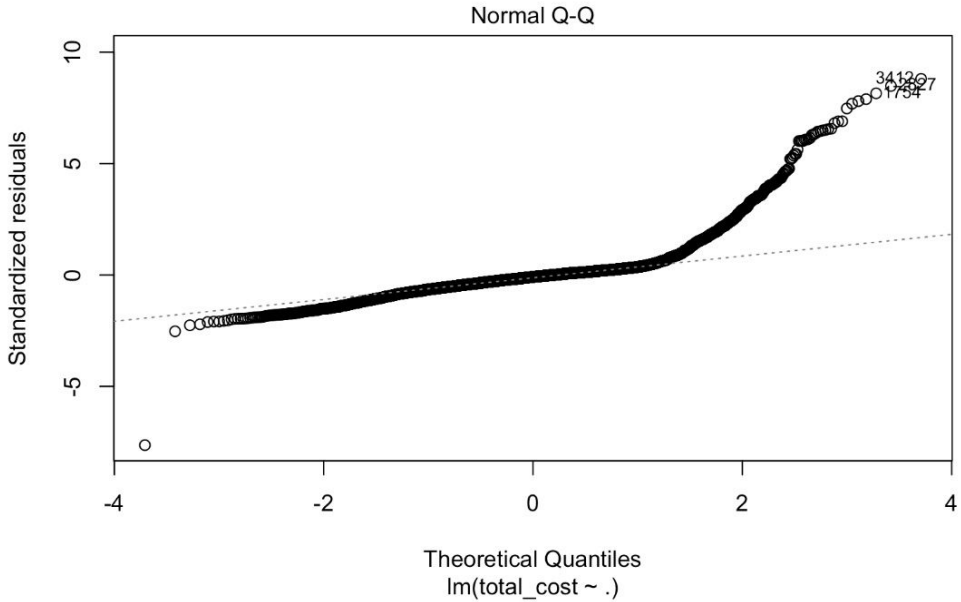
Residual term is not normally distributed & Outliers

**Observations**

- About 40% overlapped normal line
- Heavy tails with outliers

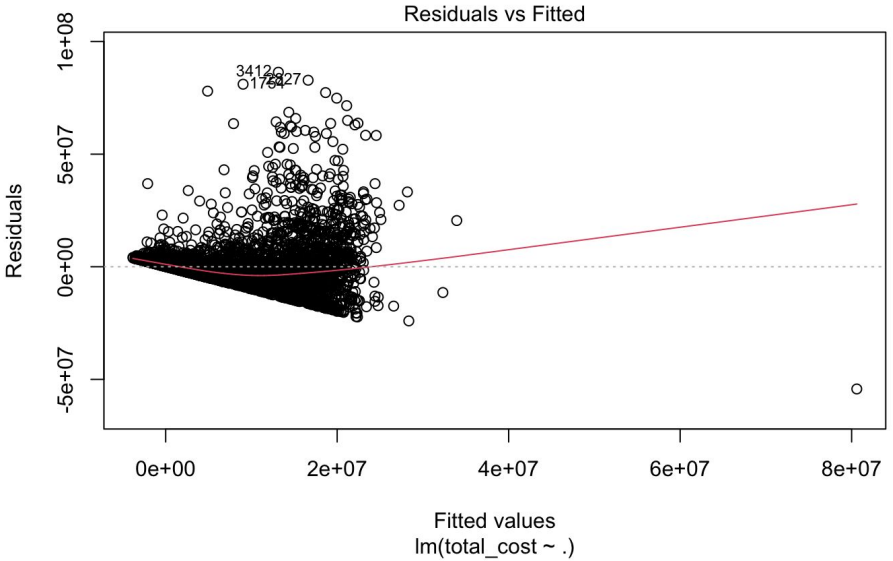
**Effects**

- Biased model
- No longer present the true line
- Tests & Confidence Intervals are approximate
- Prediction Interval is incorrect



# PRELIMINARY ANALYSIS

## Heteroskedasticity & Outliers




```
        studentized Breusch-Pagan test

data:  ols_raw
BP = 377.91, df = 38, p-value < 2.2e-16
```

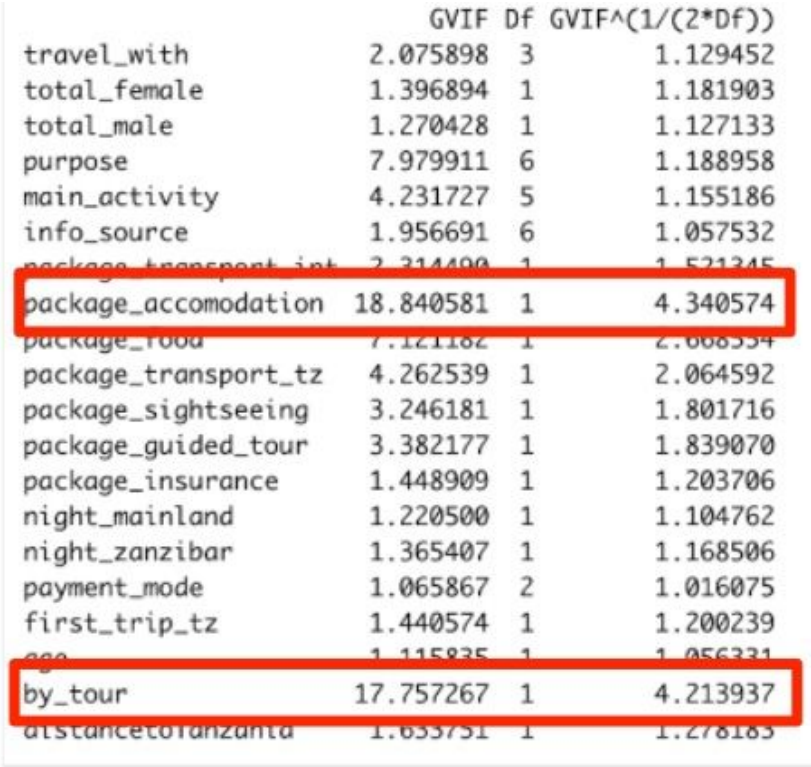
```
  Brown-Forsythe Test (alpha = 0.05)
-------------------------------------------------------------
  data : total_cost and night_mainland50

  statistic  : 7.373642
  num df     : 1
  denom df   : 60.35584
  p.value    : 0.008619533

  Result     : Difference is statistically significant.
-------------------------------------------------------------
```

# PRELIMINARY ANALYSIS

## Multicollinearity

- Approximate Chi-square Tests

| Variable 1 | Variable 2 | Chi-sq P-value |
|---|---|---|
| by_tour | package_accomodation | 0 |
| package_insurance | by_tour | 6.53E-212 |
| by_tour | package_food | 0 |
| by_tour | main_activity | 3.79E-130 |
| by_tour | package_sightseeing | 0 |
| by_tour | package_transport | 0 |

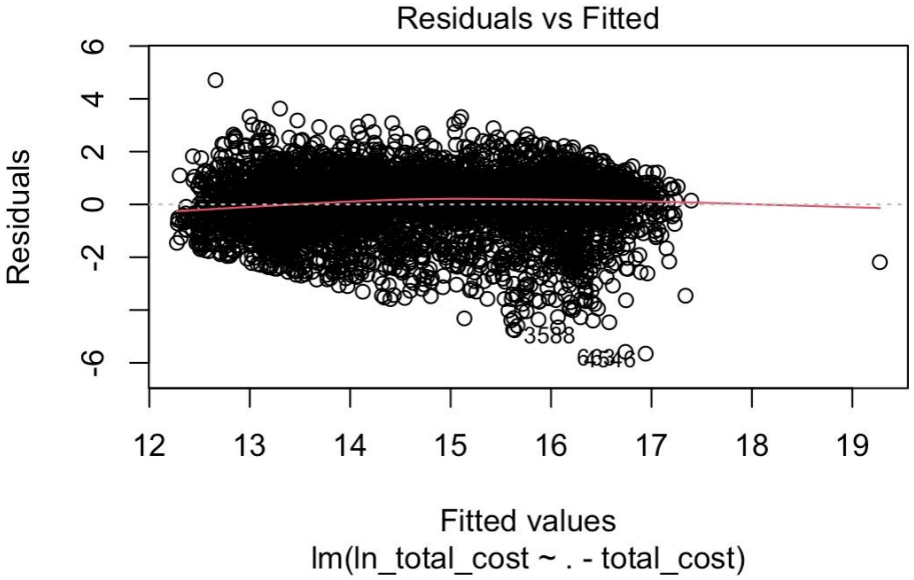


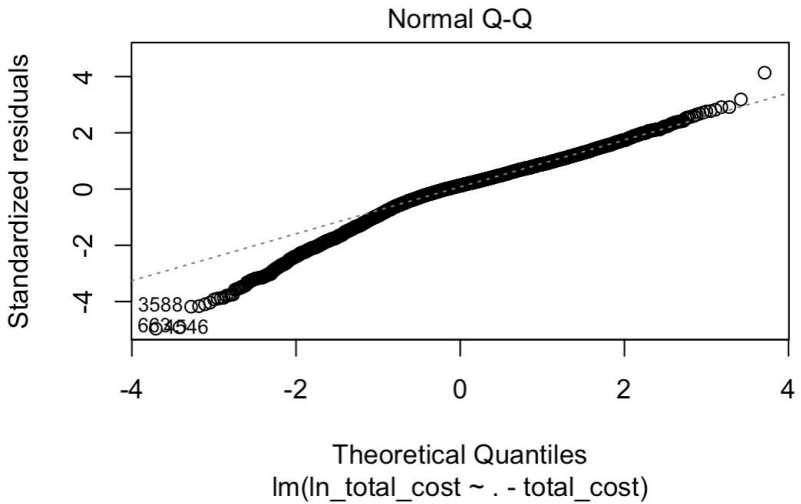
| | GVIF | Df | GVIF^(1/(2*Df)) |
|---|---|---|---|
| travel_with | 2.075898 | 3 | 1.129452 |
| total_female | 1.396894 | 1 | 1.181903 |
| total_male | 1.270428 | 1 | 1.127133 |
| purpose | 7.979911 | 6 | 1.188958 |
| main_activity | 4.231727 | 5 | 1.155186 |
| info_source | 1.956691 | 6 | 1.057532 |
| package_accomodation | 18.840581 | 1 | 4.340574 |
| package_transport_tz | 4.262539 | 1 | 2.064592 |
| package_sightseeing | 3.246181 | 1 | 1.801716 |
| package_guided_tour | 3.382177 | 1 | 1.839070 |
| package_insurance | 1.448909 | 1 | 1.203706 |
| night_mainland | 1.220500 | 1 | 1.104762 |
| night_zanzibar | 1.365407 | 1 | 1.168506 |
| payment_mode | 1.065867 | 2 | 1.016075 |
| first_trip_tz | 1.440574 | 1 | 1.200239 |
| by_tour | 17.757267 | 1 | 4.213937 |

# DATA ETL

## Remedies - Outcome Variable Transformation

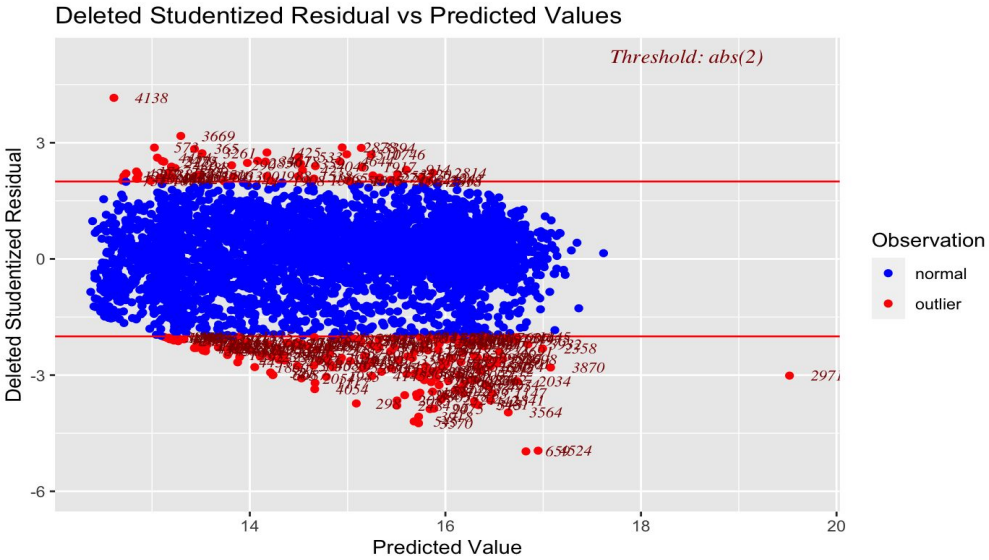
- Apply **log(total_cost)** = **ln_total_cost** (Log - Linear Elastic model)

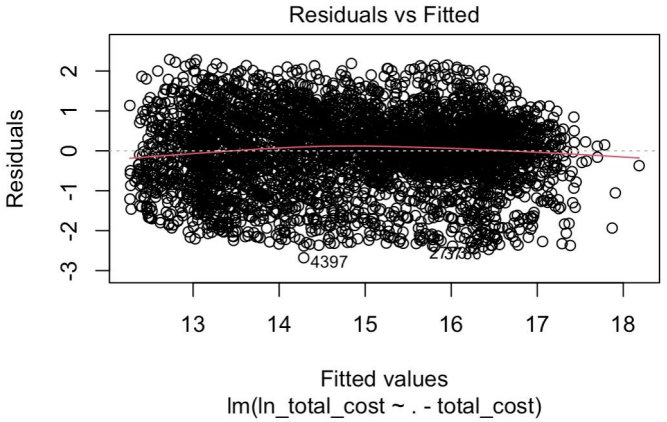


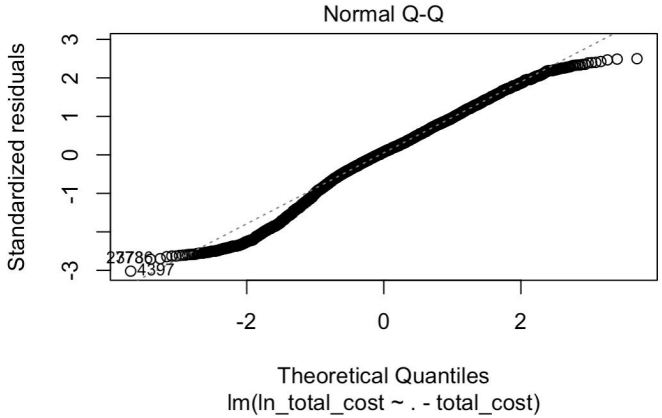



# DATA ETL

## Remedies - Outliers Removal



# DATA ETL

## Remedies - Overcome Multicollinearity by Variables Selection

**Model 1** - Full Model

| | GVIF | Df | GVIF^(1/(2*Df)) |
|---|---|---|---|
| travel_with | 2.158747 | 3 | 1.136842 |
| total_female | 1.386035 | 1 | 1.177300 |
| total_male | 1.276789 | 1 | 1.129951 |
| purpose | 8.312186 | 6 | 1.193007 |
| main_activity | 4.333972 | 5 | 1.157947 |
| info_source | 1.960815 | 6 | 1.057718 |
| package_transport_int | 2.327827 | 1 | 1.525722 |
| package_accomodation | 18.867460 | 1 | 4.343669 |
| package_food | 7.169458 | 1 | 2.677584 |
| package_transport_tz | 4.275501 | 1 | 2.067729 |
| package_sightseeing | 3.241174 | 1 | 1.800326 |
| package_guided_tour | 3.372885 | 1 | 1.836542 |
| package_insurance | 1.449504 | 1 | 1.203953 |
| night_mainland | 1.230039 | 1 | 1.109071 |
| night_zanzibar | 1.373330 | 1 | 1.171892 |
| payment_mode | 1.072967 | 2 | 1.017763 |
| first_trip_tz | 1.462391 | 1 | 1.209294 |
| age | 1.399112 | 2 | 1.087585 |
| by_tour | 17.779712 | 1 | 4.216600 |
| distancetoTanzania | 1.639418 | 1 | 1.280398 |

**Stepwise method**

**Model 2** - Reduced Model

| | GVIF | Df | GVIF^(1/(2*Df)) |
|---|---|---|---|
| travel_with | 2.105487 | 3 | 1.132119 |
| total_female | 1.385193 | 1 | 1.176942 |
| total_male | 1.272786 | 1 | 1.128178 |
| purpose | 6.662388 | 6 | 1.171213 |
| main_activity | 3.923957 | 5 | 1.146496 |
| package_transport_int | 2.116233 | 1 | 1.454728 |
| package_food | 6.298867 | 1 | 2.509754 |
| package_transport_tz | 3.921108 | 1 | 1.980179 |
| night_mainland | 1.214175 | 1 | 1.101896 |
| night_zanzibar | 1.346225 | 1 | 1.160269 |
| payment_mode | 1.063710 | 2 | 1.015561 |
| age | 1.364108 | 2 | 1.080718 |
| by_tour | 8.747095 | 1 | 2.957549 |
| distancetoTanzania | 1.591491 | 1 | 1.261543 |

# TEST STATISTIC

## Inference Goal

Examining if people at different age will spend differently on traveling.

*H0 : $B_{age} = 0$*

*HA: $B_{age} \neq 0$*

*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.*

# TEST STATISTIC

## Models' Goodness of Fit

**Model 1** - Full Model
Outcome: ln_total_cost
Predictors: 20 variables (all variables)

**Model 2 -** Reduced Model
Outcome: ln_total_cost
Predictors: 14 variables (excluding: package_accommodation, info_source, first_trip_tz, package_sightseeing, package_guided_tour, package_insurance)

**3 Criteria for Models comparison**:

- Adjusted R-squared
- AIC and BIC
- F-test on the Goodness of Fit

# TEST STATISTIC

## Models' Goodness of Fit

| 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

- **Adjusted R-sq**: Full Model (mod1) > Reduced Model (mod2)
- **AIC / BIC:** Full Model (mod1) > Reduced Model (mod2)
- **Note:** mod1 contains severe multicollinearity, while it's tolerable in mod2

# TEST STATISTIC

## Models' Goodness of Fit

- **ANOVA Breakdown (Full Model)**

| | Source <chr> | Df <dbl> | SS <dbl> | MS <dbl> | F <dbl> | P <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 |

- **ANOVA Test comparing 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
```

# MODEL TRAINING & VALIDATION

- Ordinary Least Squares Regression
- Ridge Regression
- Principle of Component Regression (PCR)

**Model Validations**

- Cross-validation using **10-fold CV** method
- Select best model based on cross-validation RMSE

| | GVIF | Df | GVIF^(1/(2*Df)) |
|---|---|---|---|
| travel_with | 2.105487 | 3 | 1.132119 |
| total_female | 1.385193 | 1 | 1.176942 |
| total_male | 1.272786 | 1 | 1.128178 |
| purpose | 6.662388 | 6 | 1.171213 |
| main_activity | 3.923957 | 5 | 1.146496 |
| package_transport_int | 2.116233 | 1 | 1.454728 |
| package_food | 6.298867 | 1 | 2.509754 |
| package_transport_tz | 3.921108 | 1 | 1.980179 |
| night_mainland | 1.214175 | 1 | 1.101896 |
| night_zanzibar | 1.346225 | 1 | 1.160269 |
| payment_mode | 1.063710 | 2 | 1.015561 |
| age | 1.364108 | 2 | 1.080718 |
| by_tour | 8.747095 | 1 | 2.957549 |
| distancetoTanzania | 1.591491 | 1 | 1.261543 |

# MODEL TRAINING & VALIDATION

## Ordinary Least Squares Regression

**Interpretation for Age**

(On average, keeping everything else constant)

- Senior spent more than Early Adult (25 - 44 years old) by 22 %
- Youth spent less than Early Adult by 20%

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

```
                                          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 TRAINING & VALIDATION

## Ridge Regression with the Best Lambda

**Interpretation for Age**

(On average, keeping everything else constant)

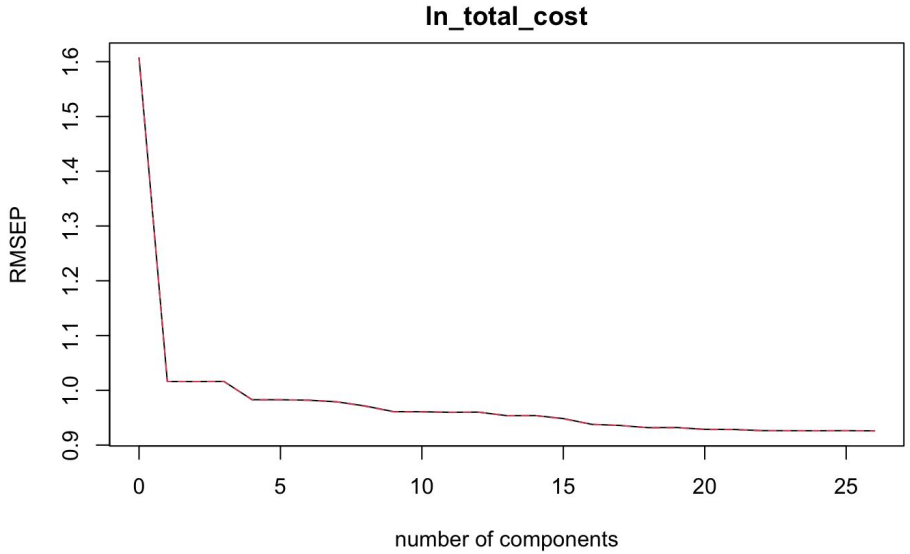
- Senior spent more than Early Adult (25 - 44 years old) by 21.77 %
- Youth spent less than Early Adult by 18.49%

| Best Ridge Lambda | Best Ridge MSE | Best Ridge RMSE |
|---|---|---|
| 0.1053521 | 0.8500239 | 0.9219674 |

| | |
|---|---|
| (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 TRAINING & VALIDATION

## Principles of Components (PCR)




```
Number of components considered: 27

VALIDATION: RMSEP
Cross-validated using 10 random segments.

       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

                27 comps
X                100.00
ln_total_cost     67.55
```

# MODEL TRAINING & VALIDATION

## Principles of Components (PCR)

**Interpretation for Age**

(On average, keeping everything else constant)

- Senior spent more than Early Adult (25 - 44 years old) by 10.54 %
- Youth spent less than Early Adult by 6.68 %

```
Number of components considered: 27

                                        24 comps  25 comps  26 comps  27 comps
VALIDATION: RMSEP                 CV      0.9228    0.9228    0.9226    0.9223
Cross-validated using 10 random segments. adjCV   0.9225    0.9225    0.9222    0.9220
```

| | 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 |

# CONCLUSIONS

## Model Selections Per Analytics Goal

### Prediction Accuracy Goal

- Rules: Select the model with the smallest cross-validation RMSE

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

# CONCLUSIONS

## Model Selections Per Analytics Goal

**Inference goal**

- Age is a significant predictor of total ecotour costs

  *“On average, Gen X-ers tend to spend more on tourism activities than any other age range does “*

- Age levels in our data vs Gen X-ers (41 - 56 years old)

  Youth (1 - 24 years old)

  Early Adult (25-44 years old)

  Late Adult (45-65 years old)

  Senior (> 65 years old)

- **The statement is not totally correct in our case!**

# CONCLUSIONS

## Model Selections Per Analytics Goal

### Interpretation Goal

- OLS Regression or any Ridge Regression with small shrinkage lambda
- 14 variables in Model 2 specifications are significant

| | |
|---|---|
| **Age:** | Senior > Earlier Adult > Youth |
| **Payment mode:** | Credit card > Cash > Others |
| **Travel_with**: | Family > Friends/Relatives > Solo > Others |

- PCR is only for prediction accuracy

# THANK YOU!

# 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. 2 135–144, Accessed from, https://academic.oup.com/tse/article/1/2/135/5618805 by American University Library user on 19 March 2022

# DATA ETL

## Remedies - Overcome Multicollinearity by Variables Selection

**Stepwise Method**

- This method will compare the significant level of explained proportion of each predictors toward the outcome variable, with upper limit of full model and lower limit of NULL model.
- The process will remove the least significant variable then fit the model again, then put back one of the removed variable to test if there's any changes in its significant.
- This loop of process will be implemented until when there is no more insignificant variable to be removed.

# MODEL TRAINING & VALIDATION

## 10-Fold Cross Validation Method

- Data will be randomly splitted into 10 folds (10 subsets)
- Everytime 9 subsets will be taken to fit a model, then the model will be validated/tested against the other 1 subset
- This loop will be repeated up to when all data points are being tested
- This model building & validation method allows us to build and validate a model at the same time, and thus, extract the best model for further model comparisons and selections.