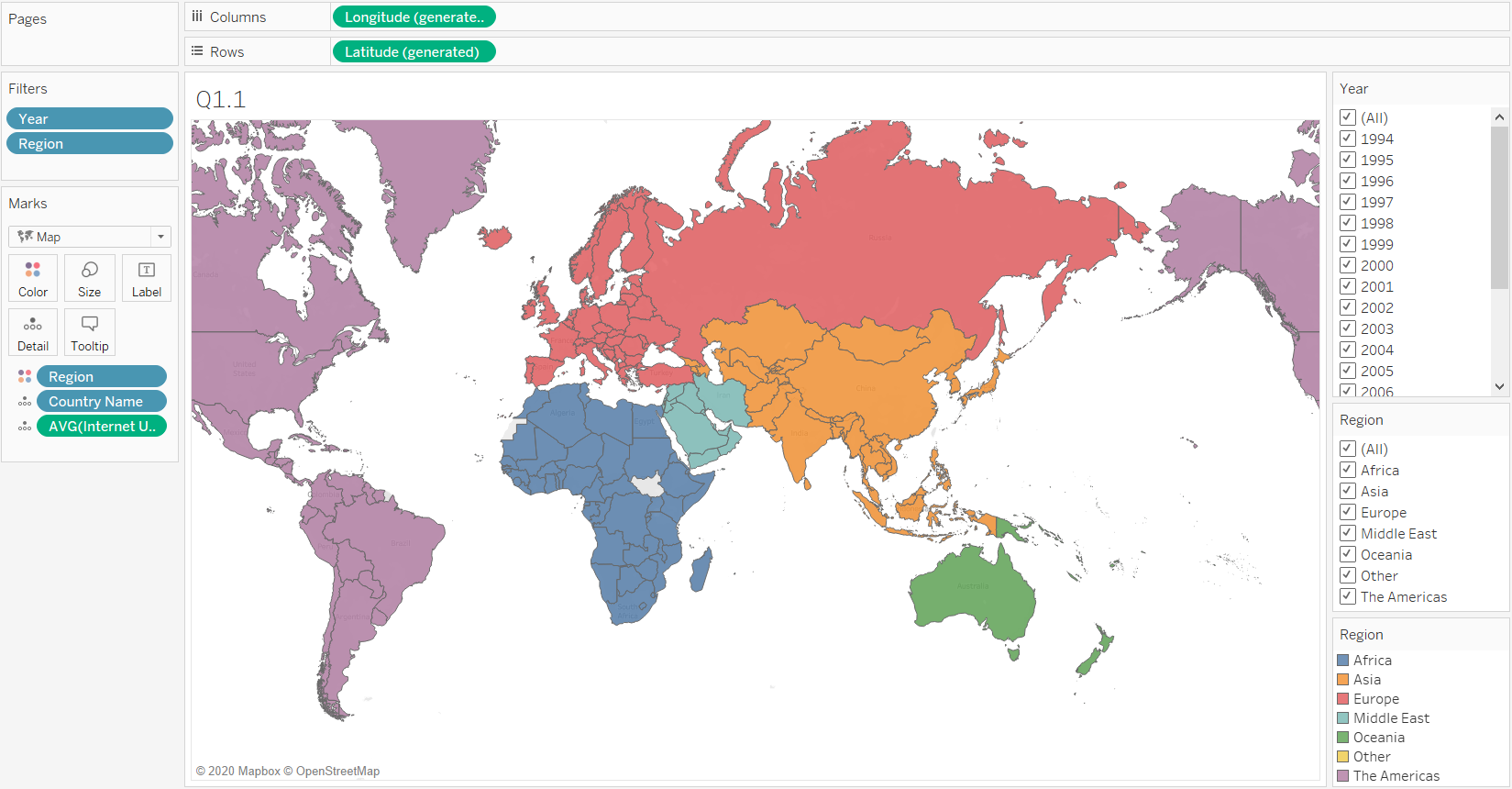
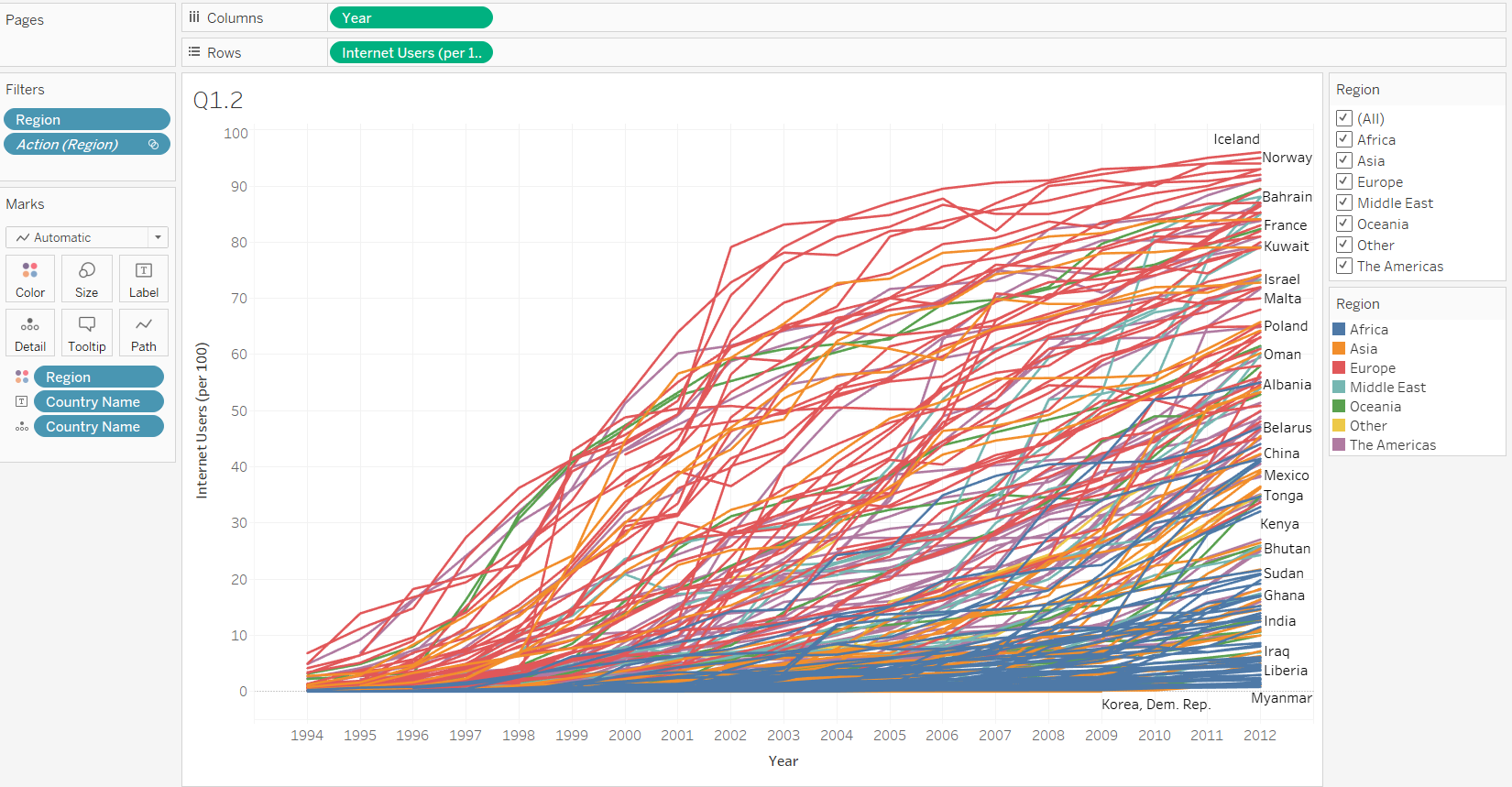
**ITM 818 Data Management and Visualization**

Homework 3: Data Visualization and Linear Regression

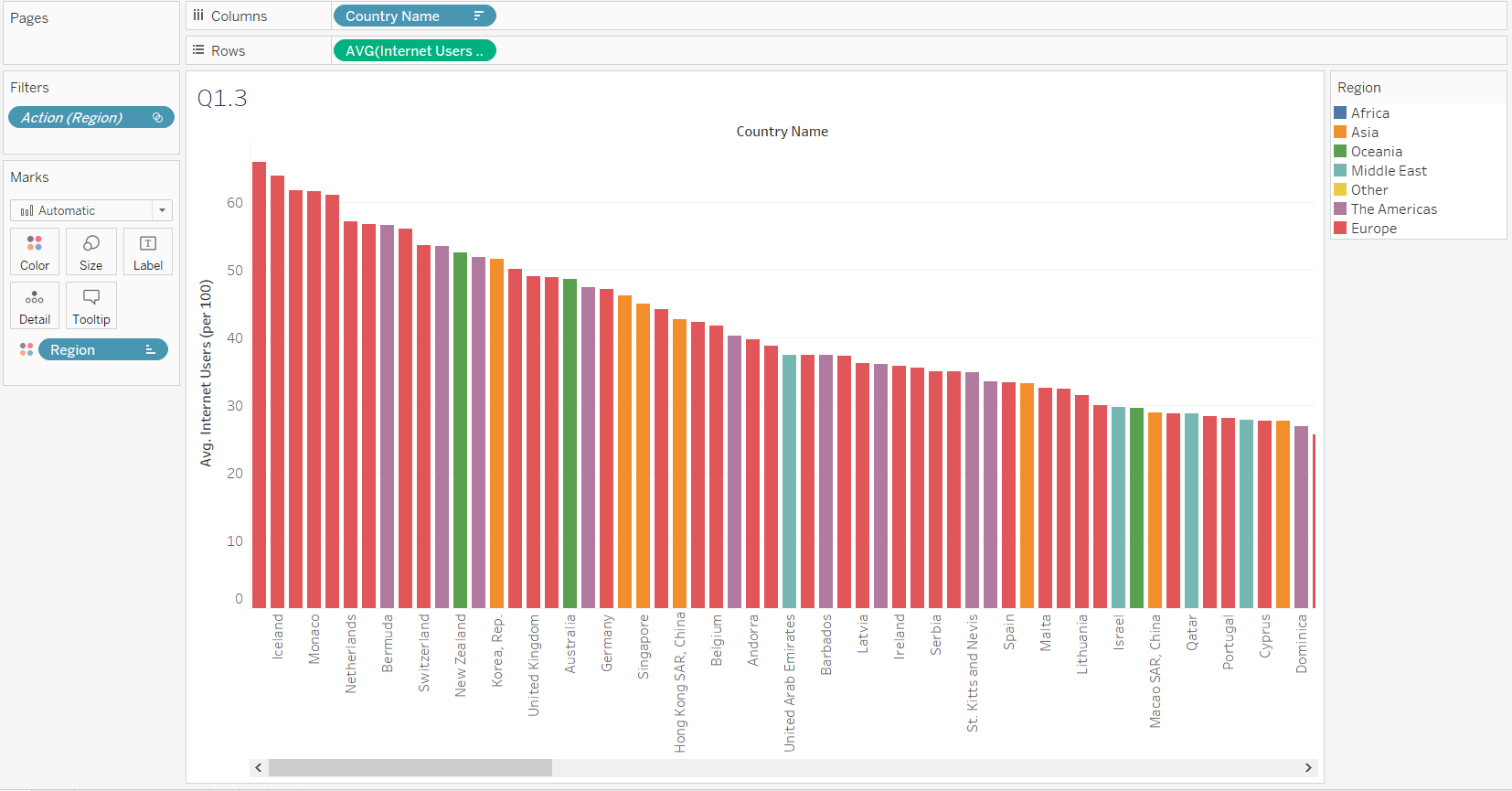
1. Please use Tableau to answer the following questions, using WorldBankInternetUsage.xlsx (20 points).  
*Rubric: major mistake if one uses the wrong visualization or required information/action is missing in the graph. Minor mistake if the visualization does not properly display the information to the readers.*

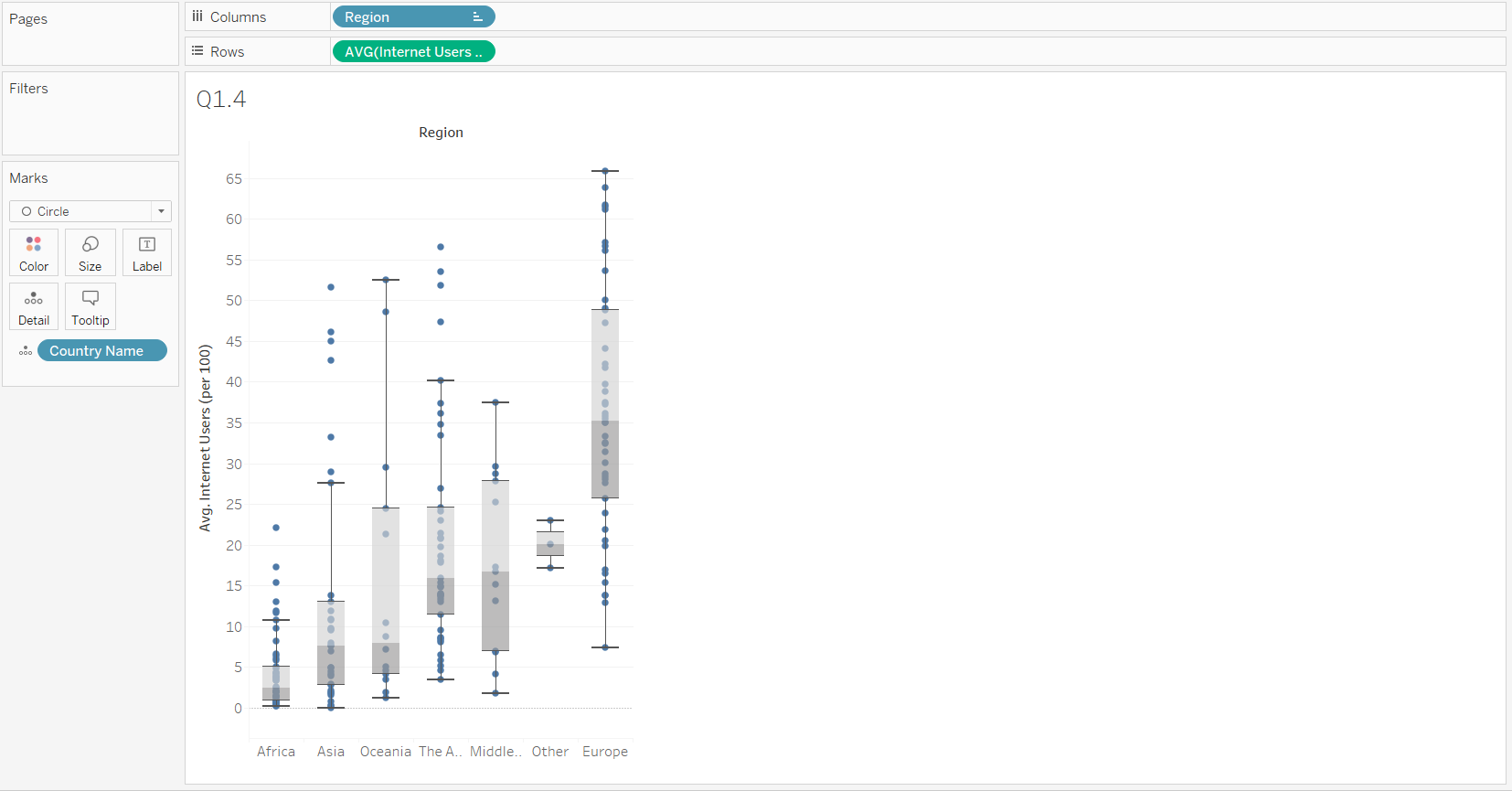
(1) Plot ***average Internet users (per 100) of each country*** in a world map over 1994-2012 data. Color the countries by region. Let the user select year and region to zoom into.

  
(2) Show timelines of Internet users (per 100) for each country. Draw one timeline for each country and color the lines by region. Let the user select regions to display.

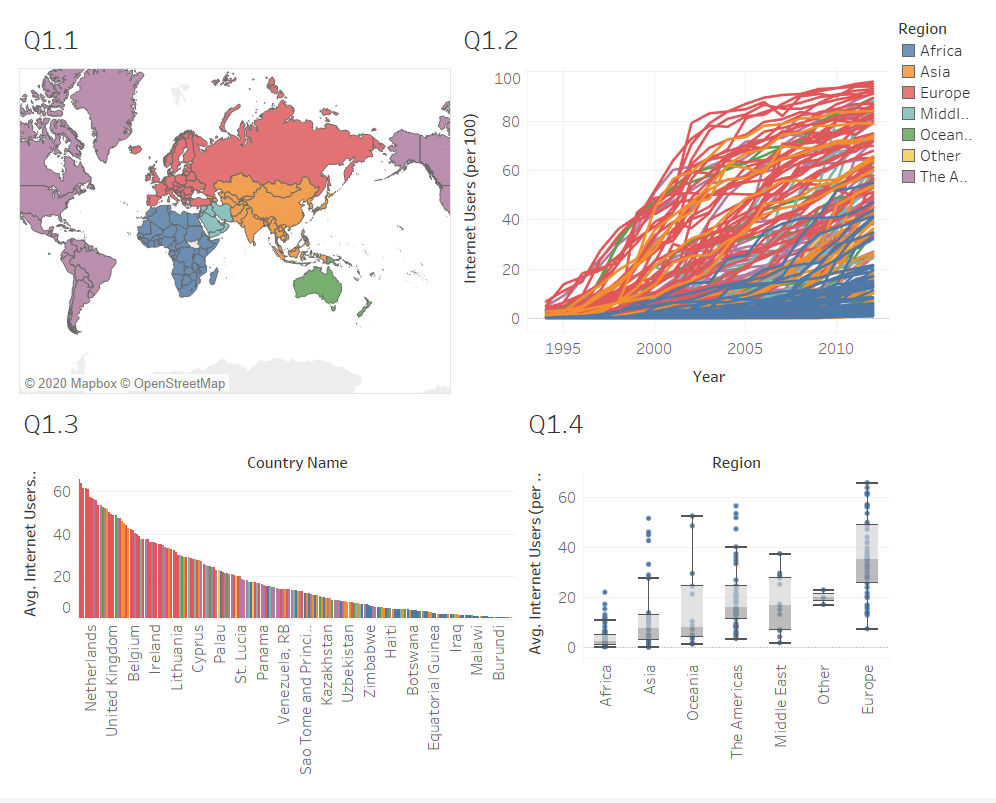


(3) For all year data (from 1994 to 2012), visualize ***average Internet users (per 100) of each country*** using a bar chart, sort the bars by Internet users (per 100). Color the countries by region.

  
(4) For all year data (from 1994 to 2012), use side-by-side boxplots to compare the distribution of ***average Internet users (per 100) of each country*** across different regions. Rank the boxplots by the median of the distribution.



(5) Create an interactive dashboard that allows the user to view all the above visuals. When the user left-clicks on a country in the map, display the timeline and rank of countries within the same region only. For example, if a user clicks France on the map, then it should display European countries in the timeline and rank chart. When the user hovers on a country on the map, highlight the timeline and rank of countries within the region.



2. Predicting Airfare on New Routes, using Airfares.csv (40 points).

The following problem takes place in the United States in the late 1990s, when many major US cities were facing issues with airport congestion, partly as a result of the 1978 deregulation of airlines. Both fares and routes were freed from regulation, and low-fare carriers such as Southwest (SW) began competing on existing routes and starting non- stop service on routes that previously lacked it. Building completely new airports is generally not feasible, but sometimes decommissioned military bases or smaller municipal airports can be reconfigured as regional or larger commercial airports. There are numerous players and interests involved in the issue (airlines, city, state and federal authorities, civic groups, the military, airport operators), and an aviation consulting firm is seeking advisory contracts with these players. The firm needs predictive models to support its consulting service. One thing the firm might want to be able to predict is fares, in the event a new airport is brought into service. The firm starts with the file Airfares.csv, which contains real data that were collected between Q3-1996 and Q2-1997. The variables in these data are listed in the following table and are believed to be important in predicting FARE. Some airport-to-airport data are available, but most data are at the city-to-city level. One question that will be of interest in the analysis is the effect that the presence or absence of Southwest has on FARE.

**Table 1. Description of Variables for Airfare Example**

|  |  |
| --- | --- |
| Variable | Description |
| S\_CODE | Starting airport’s code |
| S\_CITY | Starting city |
| E\_CODE | Ending airport’s code |
| E\_CITY | Ending city |
| COUPON | Average number of coupons (a one-coupon flight is a nonstop flight, a two-coupon flight is a one-stop flight, etc.) for that route |
| NEW | Number of new carriers entering that route between Q3-96 and Q2-97 |
| VACATION | Whether (Yes) or not (No) a vacation route |
| SW | Whether (Yes) or not (No) Southwest Airlines serves that route |
| HI | Herfindahl index: measure of market concentration of the airline industry |
| S\_INCOME | Starting city’s average personal income |
| E\_INCOME | Ending city’s average personal income |
| S\_POP | Starting city’s population |
| E\_POP | Ending city’s population |
| SLOT | Whether or not either endpoint airport is slot-controlled (this is a measure of airport congestion) |
| GATE | Whether or not either endpoint airport has gate constraints (this is another measure of airport congestion) |
| DISTANCE | Distance between two endpoint airports in miles |
| PAX | Number of passengers on that route during period of data collection |
| FARE | Average fare on that route |

1. Explore the numerical predictors and response (FARE) by creating a correlation table and examining some scatterplots between FARE and those predictors. What seems to be the best single predictor of FARE?

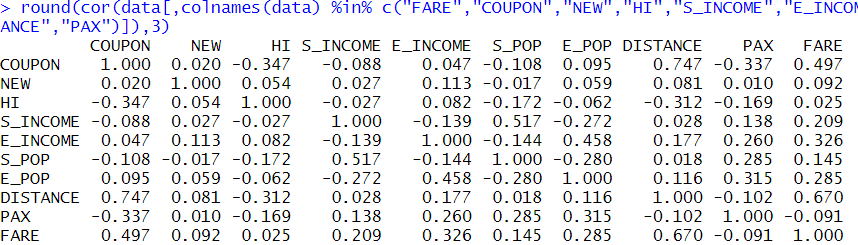
**Code**:

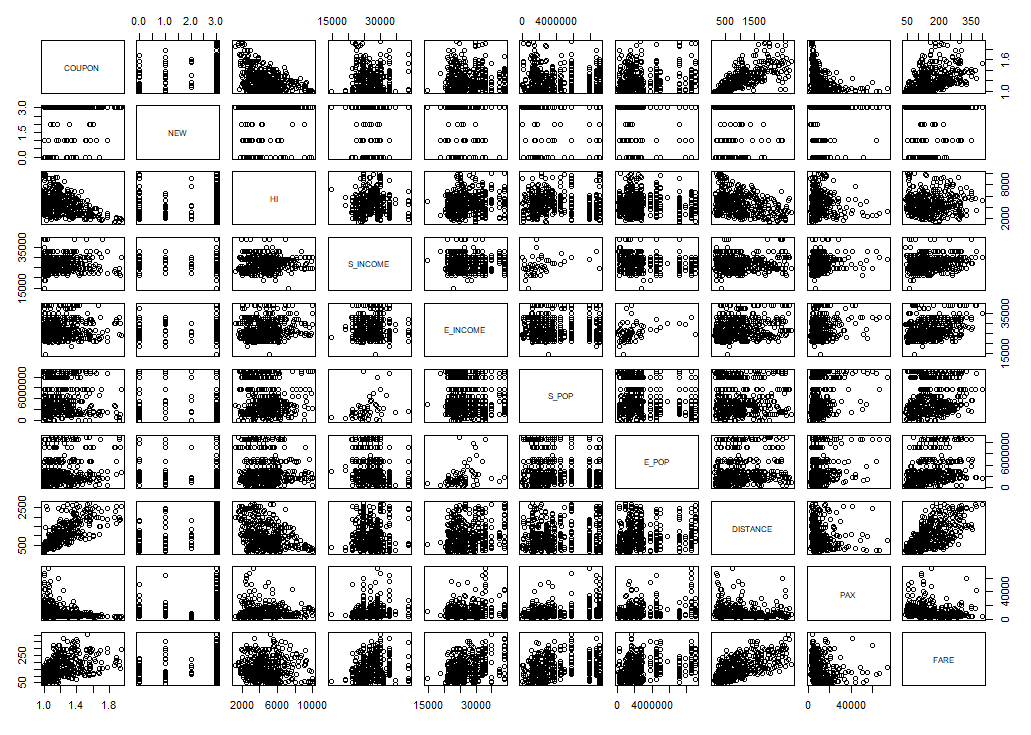
#Q2.1

round(cor(data[,colnames(data) %in% c("FARE","COUPON","NEW","HI","S\_INCOME","E\_INCOME",'S\_POP','E\_POP',"DISTANCE","PAX")]),3)

plot(data[,colnames(data) %in% c("FARE","COUPON","NEW","HI","S\_INCOME","E\_INCOME",'S\_POP','E\_POP',"DISTANCE","PAX")])

**Output**:





The Single Best predictor of fare seems to be Distance

1. Explore the categorical predictors (excluding the first four) by computing the percentage of flights in each category. Create a graphical chart to visualize the average fare in each category. Which categorical predictor seems best for predicting FARE?

**CODE:**

categorical = c('VACATION', 'SW', 'SLOT', 'GATE')

par(mfrow = c(2,2))

perc\_cat = matrix(nrow = 4, ncol = 2,

dimnames = list(c('VACATION', 'SW', 'SLOT', 'GATE'),c('Yes/Free','No/Controlled/Condensed')))

for(i in 1:(length(categorical))){

category = sort(unique(data[,categorical[i]]), decreasing = TRUE)

occ = table(data[,categorical[i]])

perc\_cat[categorical[i], 'Yes/Free'] = (occ[category[1]] \* 100)/(occ[category[1]] + occ[category[2]])

perc\_cat[categorical[i], 'No/Controlled/Condensed'] = (occ[category[2]] \* 100)/(occ[category[1]] + occ[category[2]])

avg\_fare = c()

avg\_fare[1] = mean(data$FARE[data[,categorical[i]] == category[1]])

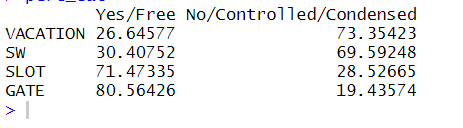
avg\_fare[2] = mean(data$FARE[data[,categorical[i]] == category[2]])

barplot(avg\_fare, names.arg = category, xlab = categorical[i], ylab = 'Avg Fare')

}

perc\_cat

**Output:**





Imp variable for predicting Fare might be SW as there is a lot of difference in the percentage plot between 2 categories.

1. Find a model for predicting the average fare on a new route:
   1. Convert categorical variables (e.g., SW) into dummy variables. Then, partition the data into training (60%) and validation (40%) sets. The model will be fit to the training data and evaluated on the validation set.
   2. Use stepwise regression to reduce the number of predictors. You can ignore the first four predictors (S\_CODE, S\_CITY, E\_CODE, E\_CITY). Report the estimated model selected.
   3. Repeat (ii) using exhaustive search instead of stepwise regression. Compare the resulting best model to the one you obtained in (ii) in terms of the predictors that are in the model.
   4. Compare the predictive accuracy of both models (ii) and (iii) using measures such as RMSE and MAPE.
   5. Using model (iii), predict the average fare on a route with the following characteristics: COUPON = 1.202, NEW = 3, VACATION = No, SW = No, HI = 4442.141, S\_INCOME = $28,760, E\_INCOME = $27,664, S\_POP = 4,557,004, E\_POP = 3,195,503, SLOT = Free, GATE = Free, PAX = 12,782, DISTANCE = 1976 miles.
   6. Predict the reduction in average fare on the route in (v) if Southwest decides to cover this route [using model (iii)].
   7. In reality, which of the factors will not be available for predicting the average fare from a new airport (i.e., before flights start operating on those routes)? Which ones can be estimated? How?
   8. Select a model that includes only factors that are available before flights begin to operate on the new route. Use an exhaustive search to find such a model.
   9. Use the model in (viii) to predict the average fare on a route with characteristics COUPON = 1.202, NEW = 3, VACATION = No, SW = No, HI = 4442.141, S\_INCOME = $28,760, E\_INCOME = $27,664, S\_ POP = 4,557,004, E\_POP = 3,195,503, SLOT = Free, GATE = Free, PAX = 12782, DISTANCE = 1976 miles.
   10. Compare the predictive accuracy of this model with model (iii). Is this model good enough, or is it worthwhile reevaluating the model once flights begin on the new route?

**RESULT:**

#2.3.1---------------------------------

data$vac\_dummy = ifelse(data$VACATION == "Yes",1,0)

data$sw\_dummy = ifelse(data$SW == "Yes",1,0)

data$slot\_dummy = ifelse(data$SLOT == "Controlled",1,0)

data$gate\_dummy = ifelse(data$GATE == "Constrained",1,0)

set.seed(1)

ti = sample(1:nrow(data),(0.6\*nrow(data)), replace = FALSE)

data1 = data[,!(names(data) %in% c("S\_CODE","S\_CITY","E\_CODE","E\_CITY","VACATION","SW","SLOT","GATE"))]

train\_a = data1[ti,]

valid\_a = data1[-ti,]

#2.3.2-------------------------------------

mod\_a = lm(FARE~ .,data=train\_a)

summary(mod\_a)

model\_aa=step(mod\_a,direction="both")

summary(model\_aa)

pred\_a=predict(model\_aa,valid\_a)

error=valid\_a$FARE-pred\_a

hist(error,breaks=30,xlab="error",main="Histogram of Error")

hist(valid\_a$FARE,breaks=30,xlab="Fare",main="Histogram of Price")

forecast::accuracy(pred\_a,valid\_a$FARE)

#2.3.3,4-------------------------------------

search=leaps::regsubsets(FARE~.,data=train\_a,nbest=1,nvmax=ncol(train\_a),method="exhaustive")

res=summary(search)

names(res)

res$which

plot(1:13,res$adjr2,type="b")

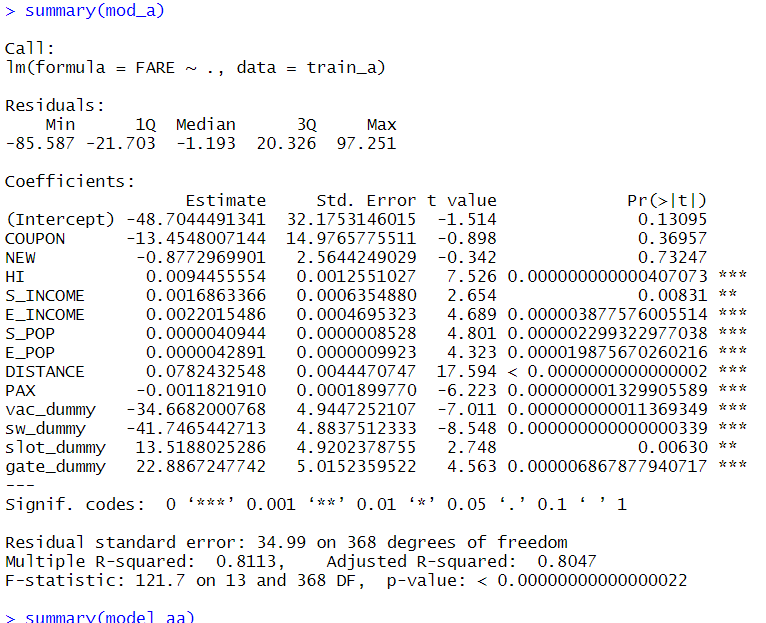
plot(1:13,res$bic,typ="b",col="red")

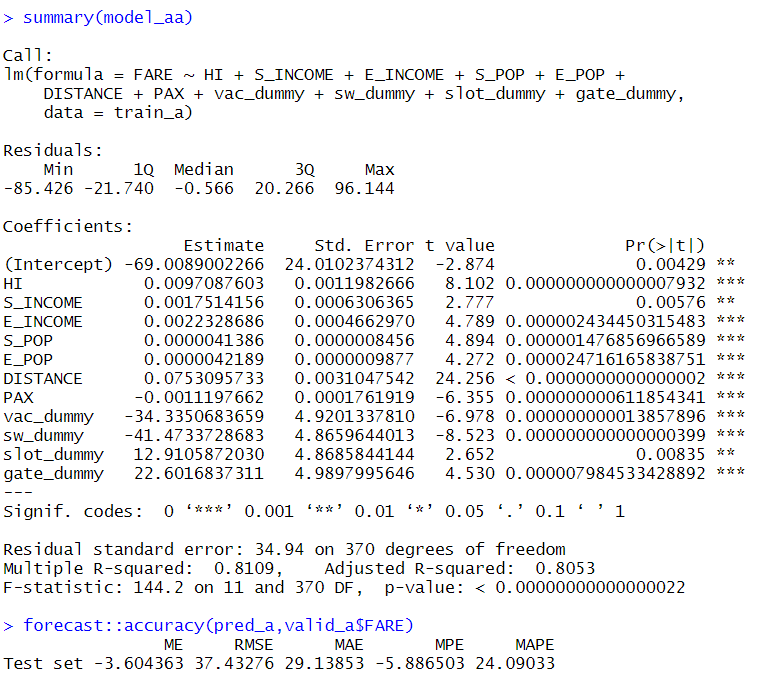
mod\_b = lm(FARE~ HI+ S\_INCOME + E\_INCOME+S\_POP+E\_POP+DISTANCE+PAX+vac\_dummy +sw\_dummy+gate\_dummy, data = train\_a)

summary(mod\_b)

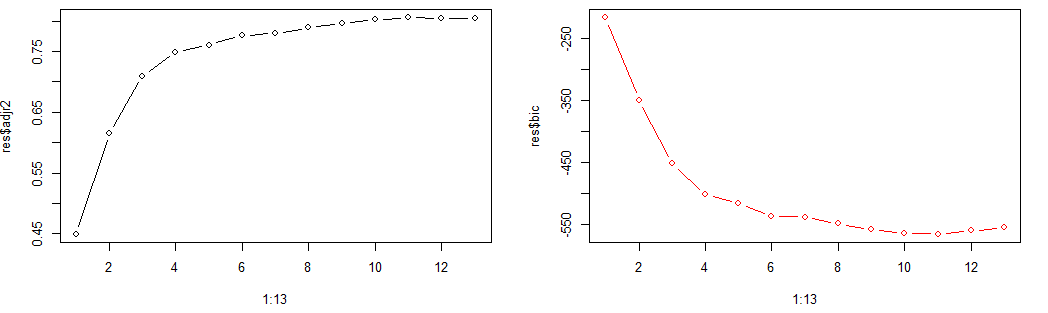
pred\_b=predict(mod\_b,valid\_a)

forecast::accuracy(pred\_b,valid\_a$FARE)

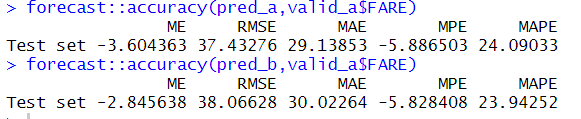




After the exhaustive search I choose number of variables = 10 as deduced from the below graph. 10 and 11 are almost similar after which the BIC increases.



Predictive Accuracy comparison:



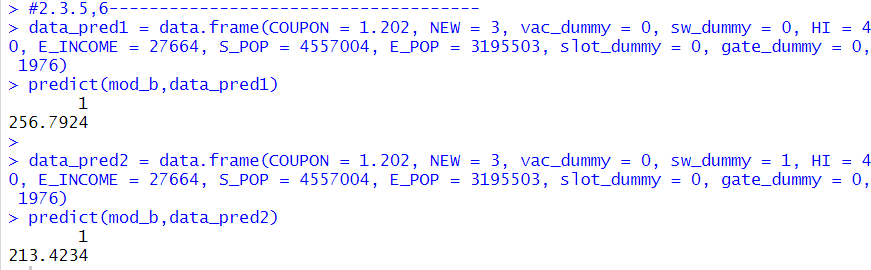
#2.3.5,6-------------------------------------

data\_pred1 = data.frame(COUPON = 1.202, NEW = 3, vac\_dummy = 0, sw\_dummy = 0, HI = 4442.141, S\_INCOME = 28760, E\_INCOME = 27664, S\_POP = 4557004, E\_POP = 3195503, slot\_dummy = 0, gate\_dummy = 0, PAX = 12782, DISTANCE = 1976)

predict(mod\_b,data\_pred1)

data\_pred2 = data.frame(COUPON = 1.202, NEW = 3, vac\_dummy = 0, sw\_dummy = 1, HI = 4442.141, S\_INCOME = 28760, E\_INCOME = 27664, S\_POP = 4557004, E\_POP = 3195503, slot\_dummy = 0, gate\_dummy = 0, PAX = 12782, DISTANCE = 1976)

predict(mod\_b,data\_pred2)



#2.3.7,8,9---------------------------------

**ANS - Following factors will be available or can be estimated to us for a new airport. S\_INCOME, E\_INCOME, S\_POP, E\_POP, DISTANCE, vac\_dummy , slot\_dummy gate\_dummy, COUPON, sw\_dummy. Most of these variables can be estimated confidently before an airport start its operations. Population and income data can be fetected from the authorities along with distance. We will know if a location is a vacation destination or not and we can estimate/confirm if Southwest is going to operate from that airport on a specific route.**

search1=leaps::regsubsets(FARE~S\_INCOME + E\_INCOME + S\_POP + E\_POP + DISTANCE + vac\_dummy + slot\_dummy + gate\_dummy+ COUPON+ sw\_dummy, data=train\_ a,nbest=1, nvmax=ncol(train\_a), method="exhaustive")

res=summary(search1)

names(res)

res$which

plot(1:10,(res$adjr2),type="b")

plot(1:10,res$bic,typ="b",col="red")

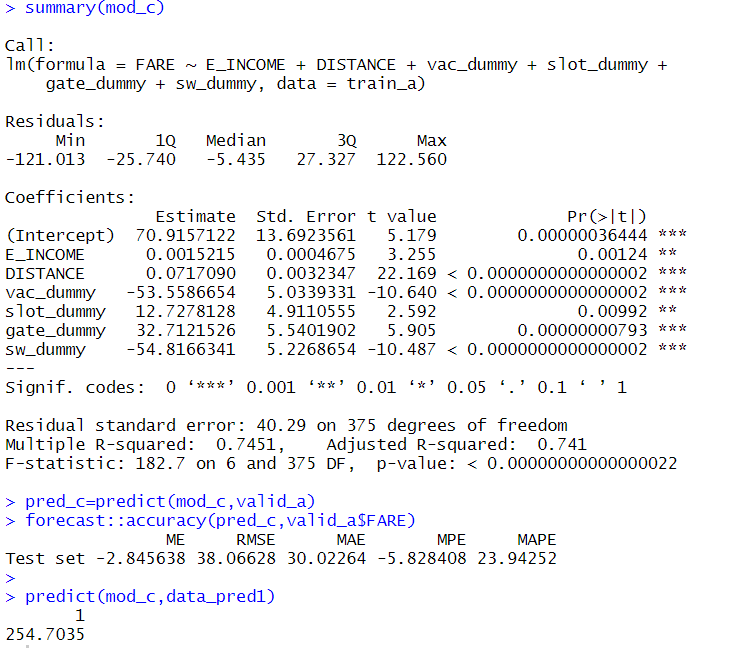
mod\_c = lm(FARE~ E\_INCOME+DISTANCE+vac\_dummy +slot\_dummy+gate\_dummy+sw\_dummy, data = train\_a)

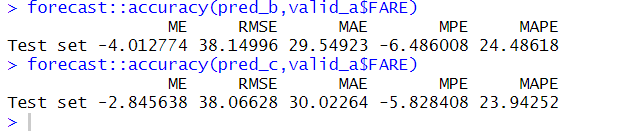
summary(mod\_c)

pred\_c=predict(mod\_c,valid\_a)

forecast::accuracy(pred\_c,valid\_a$FARE)

predict(mod\_c,data\_pred1)





This Model is good enough in predicting fares for the new airport. We can predict the fares between routes with good accuracy. Even though we have a good model I would recommend re-evaluating this after the airport operation and fights start as we can get realtime and actual data for that airport which will give us a better picture of the data and we can correct our assumptions if needed.

1. In competitive industries, a new entrant with a novel business plan can have a disruptive effect on existing firms. If a new entrant’s business model is sustainable, other players are forced to respond by changing their business practices. If the goal of the analysis was to evaluate the effect of Southwest Airlines’ presence on the airline industry rather than predicting fares on new routes, how would the analysis be different? Describe technical and conceptual aspects.

When we try to analyze the effect of Southwest airlines in the airline industry, we need to look at the routes SW operates on as well as the fares that are offered by SW vs competitive airline. We need to concentrate and findout on the difference in parameters for SW and others to understand its effect.

Technical Analysis:

Null Hypothesis: Presence of Southwest airlines does not influence airline industry. (i.e. average fares on a route with Southwest airline = average fares on a route without Southwest airline)

Alternate Hypothesis: Presence of Southwest airlines influences airline industry. (i.e. average fares on a route with Southwest airline != average fares on a route without Southwest airline)

Data requirement: We need to collect data on fares with and without the presence of Southwest airlines, keeping rest of the variables same.

Technical concepts: We will conduct a paired t-test to test our hypothesis and find out whether or not presence of Southwest airlines has any influence on airline industry.