CPM351 PROJECT PRESENTATION

"AIR TRAFFIC PASSENGER DATA" DATASET - NAIVE BAYES MODEL
PROBLEM: USING THE NAÏVE BAYES MODEL TO IDENTIFY THE ASSOCIATION BETWEEN EACH FEATURE WITH THE NUMBER OF PASSENGERS

"HOUSING PRICES DATA" DATASET - LINEAR REGRESSION MODEL
PROBLEM: USING THE LINEAR REGRESSION MODEL TO IDENTIFY THE ASSOCIATION BETWEEN THE FEATURES AND THE HOUSING PRICE.

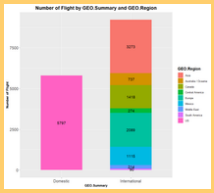
GROUP 13
PUN YI JIE (151625)
SEET ZHI NIE (151502)
TAN XIN YI (152804)

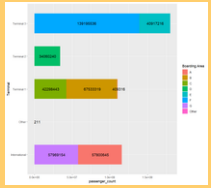
NAIVE BAYES MODEL

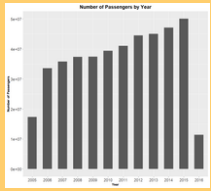
Data Preparation

```
> str(df)
'data.frame':   15007 obs. of  16 variables:
 $ Activity.Period           : int  200507 200507 200507 200507 200507 200507 200507 200507 200507 200507 ...
 $ Operating.Airline         : chr  "ATA Airlines" "ATA Airlines" "ATA Airlines" "Air Canada " ...
 $ Operating.Airline.IATA.Code: chr  "TZ" "TZ" "TZ" "AC" ...
 $ Published.Airline         : chr  "ATA Airlines" "ATA Airlines" "ATA Airlines" "Air Canada " ...
 $ Published.Airline.IATA.Code: chr  "TZ" "TZ" "TZ" "AC" ...
 $ GEO.Summary               : chr  "Domestic" "Domestic" "Domestic" "International" ...
 $ GEO.Region                : chr  "US" "US" "US" "Canada" ...
 $ Activity.Type.Code        : chr  "Deplaned" "Enplaned" "Thru / Transit" "Deplaned" ...
 $ Price.Category.Code       : chr  "Low Fare" "Low Fare" "Low Fare" "Other" ...
 $ Terminal                  : chr  "Terminal 1" "Terminal 1" "Terminal 1" "Terminal 1" ...
 $ Boarding.Area             : chr  "B" "B" "B" "B" ...
 $ Passenger.Count           : int  27271 29131 5415 35156 34090 6263 5500 12050 11638 4998 ...
 $ Adjusted.Activity.Type.Code: chr  "Deplaned" "Enplaned" "Thru / Transit * 2" "Deplaned" ...
 $ Adjusted.Passenger.Count  : int  27271 29131 10830 35156 34090 6263 5500 12050 11638 4998 ...
 $ Year                      : int  2005 2005 2005 2005 2005 2005 2005 2005 2005 2005 ...
 $ Month                     : chr  "July" "July" "July" "July" ...
```

```
> summary(df)
 Activity.Period                    Operating.Airline Operating.Airline.IATA.Code                Published.Airline
 200807 :  128   United Airlines - Pre 07/01/2013:2154   UA     :3046              United Airlines - Pre 07/01/2013:2645
 200808 :  127   Skywest Airlines                : 963   OO     : 963              United Airlines                 :1107
 201510 :  126   United Airlines                 : 892   AS     : 751              Alaska Airlines                 : 969
 201505 :  125   Alaska Airlines                 : 751   DL     : 386              Delta Air Lines                 : 803
 201603 :  125   Delta Air Lines                 : 386   AC     : 366              American Airlines               : 416
 200710 :  123   Air Canada                      : 366   VX     : 362              US Airways                      : 407
 (Other):14253   (Other)                         :9495   (Other):9133              (Other)                         :8660
 Published.Airline.IATA.Code      GEO.Summary                GEO.Region        Activity.Type.Code Price.Category.Code
 UA     :3752                 Domestic     :5797   US           :5797   Deplaned     :7071   Low Fare: 1920
 AS     : 969                 International:9210   Asia         :3273   Enplaned     :7016   Other   :13087
 DL     : 803                                      Europe       :2089   Thru / Transit: 920
 AA     : 416                                      Canada       :1418
 US     : 407                                      Mexico       :1115
 AC     : 380                                      Australia / Oceania: 737
 (Other):8280                                      (Other)      : 578
        Terminal    Boarding.Area  Passenger.Count    Adjusted.Activity.Type.Code Adjusted.Passenger.Count      Year
 International:9197   A      :5225   Min.   :    1   Deplaned            :7071   Min.   :    1           2015   :1460
 Other        :   27   G      :3992   1st Qu.:  5374   Enplaned            :7016   1st Qu.:  5496          2008   :1433
 Terminal 1   :3241   B      :1993   Median :  9210   Thru / Transit * 2: 920   Median :  9354          2007   :1409
 Terminal 2   : 324   F      :1377   Mean   : 29241                             Mean   : 29332          2009   :1393
 Terminal 3   :2218   C      :1228   3rd Qu.: 21159                             3rd Qu.: 21182          2011   :1390
                      E      : 841   Max.   :659837                             Max.   :659837          2010   :1383
                      (Other): 351                                                                      (Other):6539
     Month
 August   :1310
 July     :1303
 September:1297
 October  :1295
 January  :1268
 November :1263
 (Other)  :7271
```







Model Planning and Development

- determine the category from each features that contributes the highest conditional probability
- classify the passenger count based on the defined predictor variables

Model Technical Details

$$P(C_i|\, x_1, x_2 \ldots, x_n)\ \alpha\ \left(\prod_{j=1}^{j=n} P(x_j|\, C_i)\right) \cdot P(C_i)\ \ for\ 1 < i < k$$

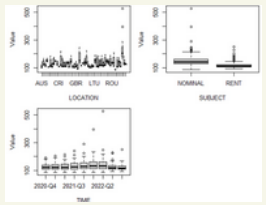
```
> cm <- confusionMatrix(test_data$Passenger,prediction)
> print(cm)
Confusion Matrix and Statistics

          Reference
Prediction  Low High
      Low  2317  298
      High  118  601

               Accuracy : 0.8752
                 95% CI : (0.8635, 0.8863)
    No Information Rate : 0.7304
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.6619

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9515
            Specificity : 0.6685
         Pos Pred Value : 0.8860
         Neg Pred Value : 0.8359
             Prevalence : 0.7304
         Detection Rate : 0.6950
   Detection Prevalence : 0.7843
      Balanced Accuracy : 0.8100

       'Positive' Class : Low
```

LINEAR REGRESSION MODEL

Data Preparation

```
> str(df1)
'data.frame':   698 obs. of  8 variables:
 $ LOCATION  : chr  "AUS" "AUS" "AUS" "AUS" ...
 $ INDICATOR : chr  "HOUSECOST" "HOUSECOST" "HOUSECOST" "HOUSECOST" ...
 $ SUBJECT   : chr  "NOMINAL" "NOMINAL" "NOMINAL" "NOMINAL" ...
 $ MEASURE   : chr  "IDX2015" "IDX2015" "IDX2015" "IDX2015" ...
 $ FREQUENCY : chr  "Q" "Q" "Q" "Q" ...
 $ TIME      : chr  "2020-Q4" "2021-Q1" "2021-Q2" "2021-Q3" ...
 $ Value     : num  116 123 131 138 144 ...
 $ Flag.Codes: logi  NA NA NA NA NA NA ...
```

```
> summary(new_df1)
    LOCATION        SUBJECT          TIME          Value
 CAN    : 17   NOMINAL:343   2020-Q4: 90   Min.   : 84.96
 CHE    : 17   RENT   :355   2021-Q1: 90   1st Qu.:110.12
 GBR    : 17                 2021-Q2: 90   Median :124.53
 IRL    : 17                 2021-Q3: 90   Mean   :132.47
 ISL    : 17                 2021-Q4: 89   3rd Qu.:146.53
 NOR    : 17                 2022-Q1: 86   Max.   :528.10
 (Other):596                 (Other):163
```



Model Planning and Development

- investigate the relationship between the continuous and categorical variables of our dataset
- categorical data as the predictor allows us to find out the average value of each categorical variable on the continuous variable

Model Technical Details

```
Call:
lm(formula = Value ~ LOCATION + SUBJECT + TIME, data = new_df1)

Residuals:
     Min       1Q   Median       3Q      Max
-55.945  -7.466   -0.985    6.479  259.507

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   127.8642     5.0872  25.135  < 2e-16 ***
LOCATIONAUT    20.0825     6.3861   3.145 0.001740 **
LOCATIONBEL     1.6426     6.3861   0.257 0.797100
LOCATIONBGR    16.0042     8.0592   1.986 0.047479 *
LOCATIONBRA   -32.2010     7.7342  -4.163 3.56e-05 ***
LOCATIONCAN    18.0139     6.3040   2.861 0.004365 **
LOCATIONCHE    -3.6278     6.3040  -0.575 0.565170
LOCATIONCHL    27.1900     6.3861   4.258 2.38e-05 ***
LOCATIONCHN    16.4013     9.0372   1.815 0.070014 .
LOCATIONCOL    15.4856     6.4794   2.390 0.017138 *
LOCATIONCRI    14.7221     9.0174   1.633 0.103040
LOCATIONCZE    33.0316     6.3861   5.172 3.10e-07 ***
LOCATIONDEU    13.2443     6.3861   2.074 0.038487 *
LOCATIONDNK     4.6688     6.3861   0.731 0.464994
LOCATIONEA      3.8195     6.3861   0.598 0.549994
LOCATIONEA17   -0.3276     8.0592  -0.041 0.967584
LOCATIONESP     1.3570     6.3861   0.212 0.831789
LOCATIONEST    32.9525     6.3861   5.160 3.30e-07 ***
LOCATIONFIN    -5.5882     6.3861  -0.875 0.381878
LOCATIONFRA    -4.1556     6.3861  -0.651 0.515458
LOCATIONGBR     2.6102     6.3040   0.414 0.678979
LOCATIONGRC   -12.1193     6.3861  -1.898 0.058180 .
LOCATIONHUN    54.6766     6.3861   8.562  < 2e-16 ***
LOCATIONIDN   -21.0943     8.0592  -2.617 0.009069 **
LOCATIONIND    -3.4314     8.0592  -0.426 0.670413
LOCATIONIRL    18.9528     6.3040   3.006 0.002747 **
LOCATIONISL    39.5771     6.3040   6.278 6.33e-10 ***
LOCATIONISR     1.3746     6.3861   0.215 0.829639
LOCATIONITA   -13.9480     6.3861  -2.184 0.029317 *
LOCATIONJPN   -10.8175     7.0120  -1.543 0.123394
LOCATIONKOR    -5.5315     6.3861  -0.866 0.386717
LOCATIONLTU    36.3800     6.3861   5.697 1.86e-08 ***
LOCATIONLUX    20.4175     6.3861   3.197 0.001456 **
LOCATIONLVA    16.3483     6.3861   2.560 0.010697 *
LOCATIONMEX    17.8654     6.4794   2.757 0.005995 **
LOCATIONNLD    21.2772     6.3861   3.332 0.000913 ***
LOCATIONNOR     5.0741     6.3040   0.805 0.421177
LOCATIONNZL    28.9536     6.5828   4.398 1.28e-05 ***
LOCATIONOECD   12.0511     8.0592   1.495 0.135322
LOCATIONPOL    23.4634     6.3861   3.674 0.000259 ***
LOCATIONPRT    22.6753     6.3861   3.551 0.000412 ***
LOCATIONROU    -1.2927     8.0592  -0.160 0.872614
LOCATIONRUS    29.4612     6.8479   4.302 1.96e-05 ***
LOCATIONSAU   -50.7253     8.0592  -6.294 5.74e-10 ***
LOCATIONSVK    12.1022     6.3861   1.895 0.058533 .
LOCATIONSVN    23.6596     6.3861   3.705 0.000230 ***
LOCATIONSWE     3.3281     6.3861   0.521 0.602446
LOCATIONTUR   120.6085     6.3861  18.886  < 2e-16 ***
LOCATIONUSA    24.3097     6.3861   3.807 0.000154 ***
LOCATIONZAF     7.5669     6.3040   1.200 0.230457
SUBJECTRENT   -37.4931     1.4296 -26.226  < 2e-16 ***
TIME2021-Q1     2.2833     2.5472   0.896 0.370376
TIME2021-Q2     4.8433     2.5472   1.901 0.057690 .
TIME2021-Q3     7.8367     2.5472   3.077 0.002183 **
TIME2021-Q4    11.1708     2.5550   4.372 1.44e-05 ***
TIME2022-Q1    15.3775     2.5807   5.959 4.20e-09 ***
TIME2022-Q2    20.1199     2.6251   7.664 6.70e-14 ***
TIME2022-Q3    16.3769     3.1568   5.188 2.86e-07 ***
TIME2022-Q4    18.0081     3.4623   5.201 2.67e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 17.09 on 639 degrees of freedom
Multiple R-squared:  0.7583,    Adjusted R-squared:  0.7364
F-statistic: 34.57 on 58 and 639 DF,  p-value: < 2.2e-16
```

FINDINGS:

Naïve Bayes Model	Linear Regression Model
Only US provided domestic flights while Asia have the most international flights.	The features that give higher housing price is the location "TUR", subject "NOMINAL" and during the time "2022-Q2".
There is an increasing number of passengers from 2005 to 2016.	The housing price increasing from "2020-Q4" until "2022-Q2", decrease on "2022-Q3" and then increase on "2022-Q4".
The data is from one of the airports in the US.	

RECOMMENDATIONS:

-The air traffic control might have to allocate some flights using terminal 3 or the international terminal to use terminal 2 to increase the air traffic efficiency.
-The housing price can always be compared using linear regression to ensure a suitable and reasonable price with the features provided.