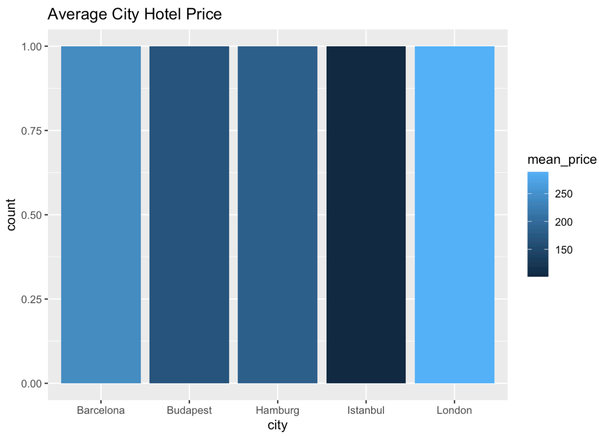
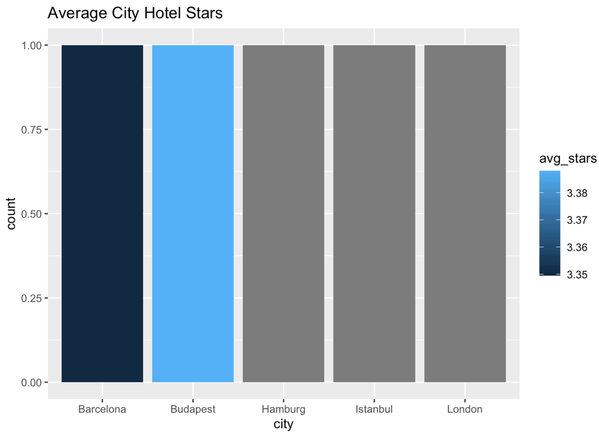
da2 a2

Abigail Chen

Introduction

For the DA2 assignment2 I used the data taken from the course material about hotels in Europe. The data can be accessed in these links **features data**, and **prices data**. I chose 5 different cities to compare, Barcelona, Budapest, Hamburg, Istanbul and London.





From the above graphs we can see that London has the highest hotel mean price, followed by Barcelona compared to the other chosen cities. However, in terms of the average stars in the cities hotels London shares a top position with both Hamburg and Istanbul.

Linear Probability Model

```
##
## Call:
## lm(formula = rating ~ distance + stars, data = hotel_five)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.91619 -0.41841  0.08876  0.34301  1.15360
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.298163   0.012351  -24.14  < 2e-16 ***
## distance    -0.008248   0.001071   -7.70 1.42e-14 ***
## stars        0.243530   0.003316   73.43  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4495 on 23728 degrees of freedom
##   (7457 observations deleted due to missingness)
## Multiple R-squared:  0.188,  Adjusted R-squared:  0.1879
## F-statistic:  2747 on 2 and 23728 DF,  p-value: < 2.2e-16
```

From the regression above we can observe that the two variables distance and stars do affect the overall rating of a hotel. It can however be noted that an increase in distance does reduce the overall rating by 0.8%.

Logit Model

```
##
## Call:
## glm(formula = rating ~ distance + stars, family = binomial(link = "logit"),
##     data = hotel_five)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0674  -1.0085   0.5064   0.8783   2.5338
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.90946    0.07124  -54.87   <2e-16 ***
## distance    -0.03722    0.00522   -7.13    1e-12 ***
## stars        1.18715    0.01964   60.45   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 32785  on 23730  degrees of freedom
## Residual deviance: 27896  on 23728  degrees of freedom
##   (7457 observations deleted due to missingness)
## AIC: 27902
##
## Number of Fisher Scoring iterations: 3
```

From the logit model an increase in distance does reduce the rating by 3.7% which is contrary to an increase in the ratings at 11.9% increase in the number of stars.

Probit Model

```
##
## Call:
## glm(formula = rating ~ distance + stars, family = binomial(link = "probit"),
##     data = hotel_five)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0862  -1.0219   0.4972   0.8850   2.6595
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.315131   0.040602 -57.021  < 2e-16 ***
## distance    -0.023776   0.003159  -7.528 5.17e-14 ***
## stars        0.706561   0.011096  63.676  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 32785  on 23730  degrees of freedom
## Residual deviance: 27928  on 23728  degrees of freedom
##   (7457 observations deleted due to missingness)
## AIC: 27934
##
## Number of Fisher Scoring iterations: 4
```