Silas\_Analysis

Max Kuttner

25 3 2020

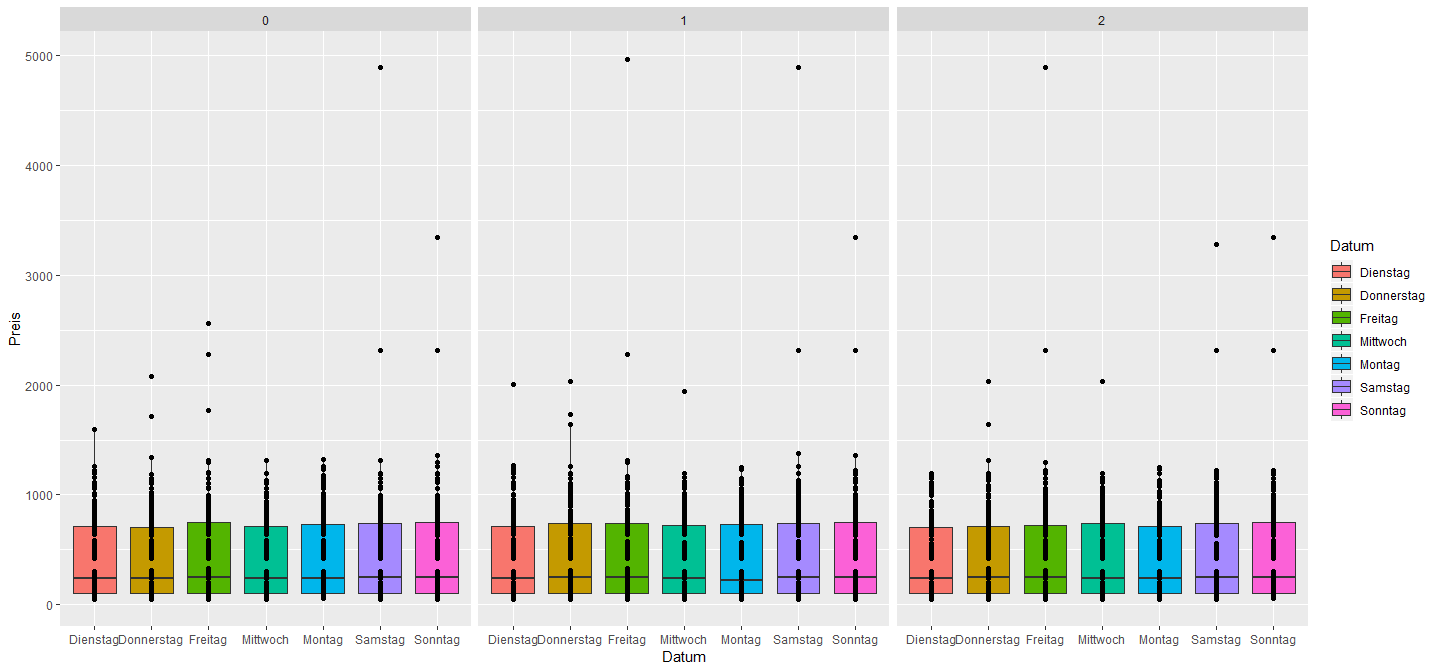
## Hypothese 1

**[Erfolgt die Buchung von Flugtickets Dienstagabends, kann statistisch gesehen das preiswerteste Offert erzielt werden.]**

### Deskreptive Statistik:

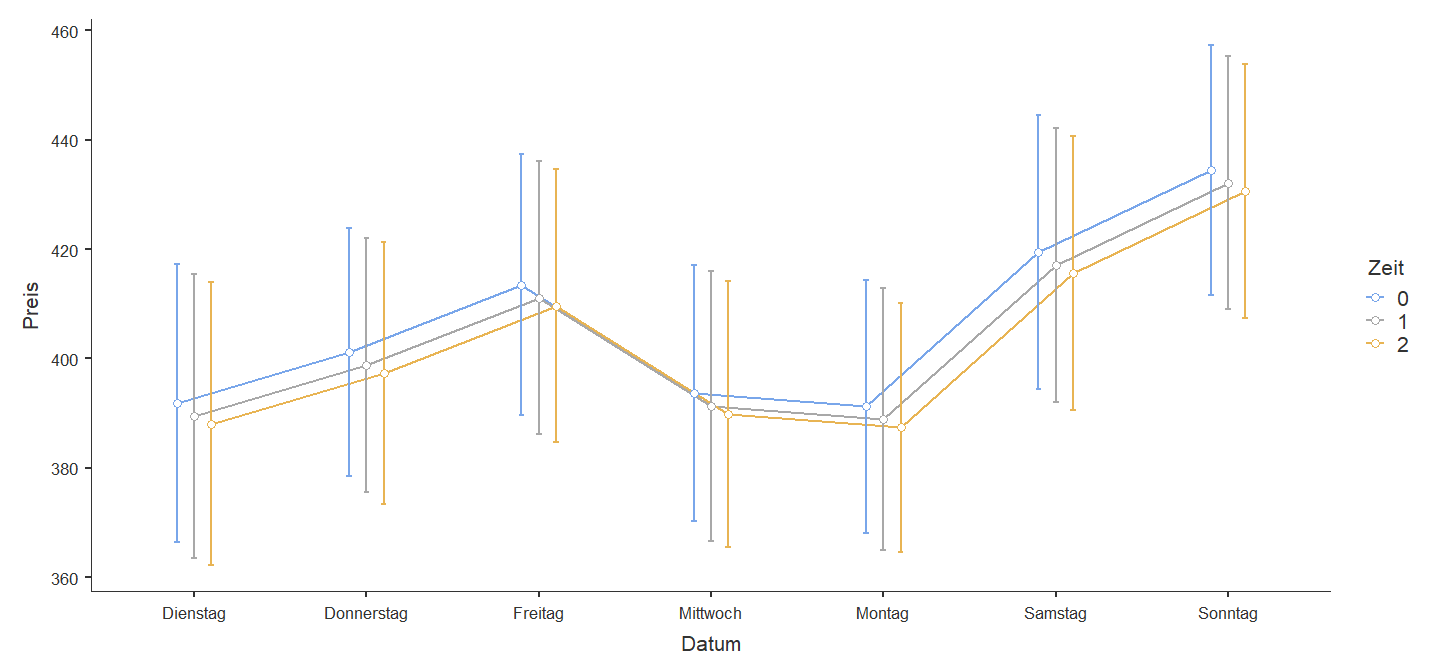
##   
## DESCRIPTIVES  
##   
## Descriptives   
## ------------------------------------------------   
## Preis Datum Zeit   
## ------------------------------------------------   
## N 8652 8652 8652   
## Missing 0 0 0   
## Mean 405   
## Median 248   
## Standard deviation 380   
## Minimum 48.0   
## Maximum 4972   
## ------------------------------------------------   
##   
##   
## FREQUENCIES  
##   
## Frequencies of Datum   
## ------------------------------------------------------   
## Levels Counts % of Total Cumulative %   
## ------------------------------------------------------   
## Dienstag 1043 12.1 12.1   
## Donnerstag 1358 15.7 27.8   
## Freitag 1176 13.6 41.3   
## Mittwoch 1225 14.2 55.5   
## Montag 1351 15.6 71.1   
## Samstag 1113 12.9 84.0   
## Sonntag 1386 16.0 100.0   
## ------------------------------------------------------   
##   
##   
## Frequencies of Zeit   
## --------------------------------------------------   
## Levels Counts % of Total Cumulative %   
## --------------------------------------------------   
## 0 3164 36.6 36.6   
## 1 2709 31.3 67.9   
## 2 2779 32.1 100.0   
## --------------------------------------------------

### Verteilung - BoxPlot



### Mehrfakorielle ANOVA:

##   
## ANOVA  
##   
## ANOVA   
## ----------------------------------------------------------------------------------   
## Sum of Squares df Mean Square F p <U+03B7>²p   
## ----------------------------------------------------------------------------------   
## Datum 2100920 6 350153 2.4252 0.024 0.002   
## Zeit 21751 2 10876 0.0753 0.927 0.000   
## Residuals 1.25e+9 8643 144379   
## ----------------------------------------------------------------------------------   
##   
##   
## ESTIMATED MARGINAL MEANS  
##   
## DATUM:ZEIT  
##   
## Estimated Marginal Means - Datum:Zeit   
## --------------------------------------------------------   
## Zeit Datum Mean SE Lower Upper   
## --------------------------------------------------------   
## 0 Dienstag 392 13.0 366 417   
## Donnerstag 401 11.5 379 424   
## Freitag 413 12.2 390 437   
## Mittwoch 394 12.0 370 417   
## Montag 391 11.8 368 414   
## Samstag 419 12.8 394 445   
## Sonntag 434 11.6 412 457   
## 1 Dienstag 390 13.2 364 415   
## Donnerstag 399 11.8 376 422   
## Freitag 411 12.7 386 436   
## Mittwoch 391 12.6 367 416   
## Montag 389 12.2 365 413   
## Samstag 417 12.8 392 442   
## Sonntag 432 11.8 409 455   
## 2 Dienstag 388 13.2 362 414   
## Donnerstag 397 12.2 373 421   
## Freitag 410 12.7 385 435   
## Mittwoch 390 12.4 366 414   
## Montag 387 11.6 365 410   
## Samstag 416 12.8 391 441   
## Sonntag 431 11.8 407 454   
## --------------------------------------------------------



Estimated marginal means korrigiert Missverhältnisse aus unterschiedlich großen Sample-Größen für einzelne Tage. Somit wird jeder/jede Tag/Uhrzeit gleich gewertet. Wie oft jeder einzelne Tag gemessen wurde bzw. im Datensatz vorkommt, ist in der deskriptiven Statistik unter **FREQUENCIES** zu sehen. Für mehr Infos zum EMM: <https://cran.r-project.org/web/packages/emmeans/vignettes/basics.html>

Im folgenden werden Tage und Uhrzeiten nach ihrem mean (also **Preis**) angeordnet.

## # A tibble: 21 x 6  
## Zeit Datum mean se lower upper  
## <fct> <fct> <dbl> <dbl> <dbl> <dbl>  
## 1 2 Montag 387. 11.6 365. 410.  
## 2 2 Dienstag 388. 13.2 362. 414.  
## 3 1 Montag 389. 12.2 365. 413.  
## 4 1 Dienstag 390. 13.2 364. 415.  
## 5 2 Mittwoch 390. 12.4 366. 414.  
## 6 0 Montag 391. 11.8 368. 414.  
## 7 1 Mittwoch 391. 12.6 367. 416.  
## 8 0 Dienstag 392. 13.0 366. 417.  
## 9 0 Mittwoch 394. 12.0 370. 417.  
## 10 2 Donnerstag 397. 12.2 373. 421.  
## # ... with 11 more rows

### Interpretation - H1

Wir sehen durch die ANOVA, dass die Variable **Datum** einen signifikanten Effekt auf den **Preis** hat. Das liest man daraus, dass der p-Wert kleiner als 0.05 ist bei 95% Signifikanzniveau. Das heißt es besteht Grund zur Annahme, dass der **Preis** vom Tag abhängt. **Zeit** hingegen ist nicht signifikant, da der p-Wert deutlich über 0.05 liegt. Das heiß die Nullhyptohese wird in diesem Fall beibehalten. Das heißt es gibt keine signifkanten Unterschiede durch die Variable Zeit.

Also folgt, dass es egal ist zu welchem Zeitpunkt man ein Ticket kaufen möchte. Wichtig ist nur, dass Montag am Abend nach Estimated marginal mean der beste Zeitpunkt ist ein Ticket zu kaufen.

## Hypothese 2

**[Je spontaner und kurzfristiger die Kaufentscheidung getroffen wird, desto höher ist der offerierte Preis einer Airline.]**

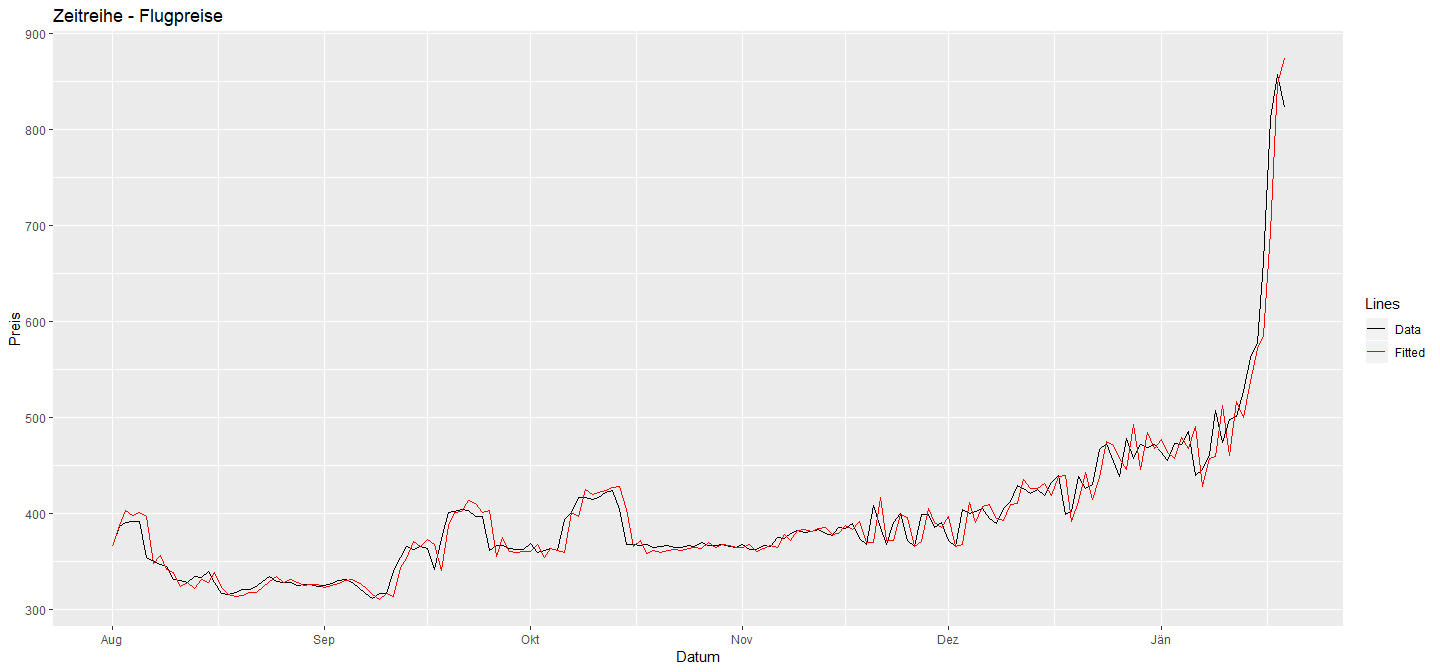
## Warning in value[[3L]](cond): The chosen test encountered an error, so no  
## seasonal differencing is selected. Check the time series data.

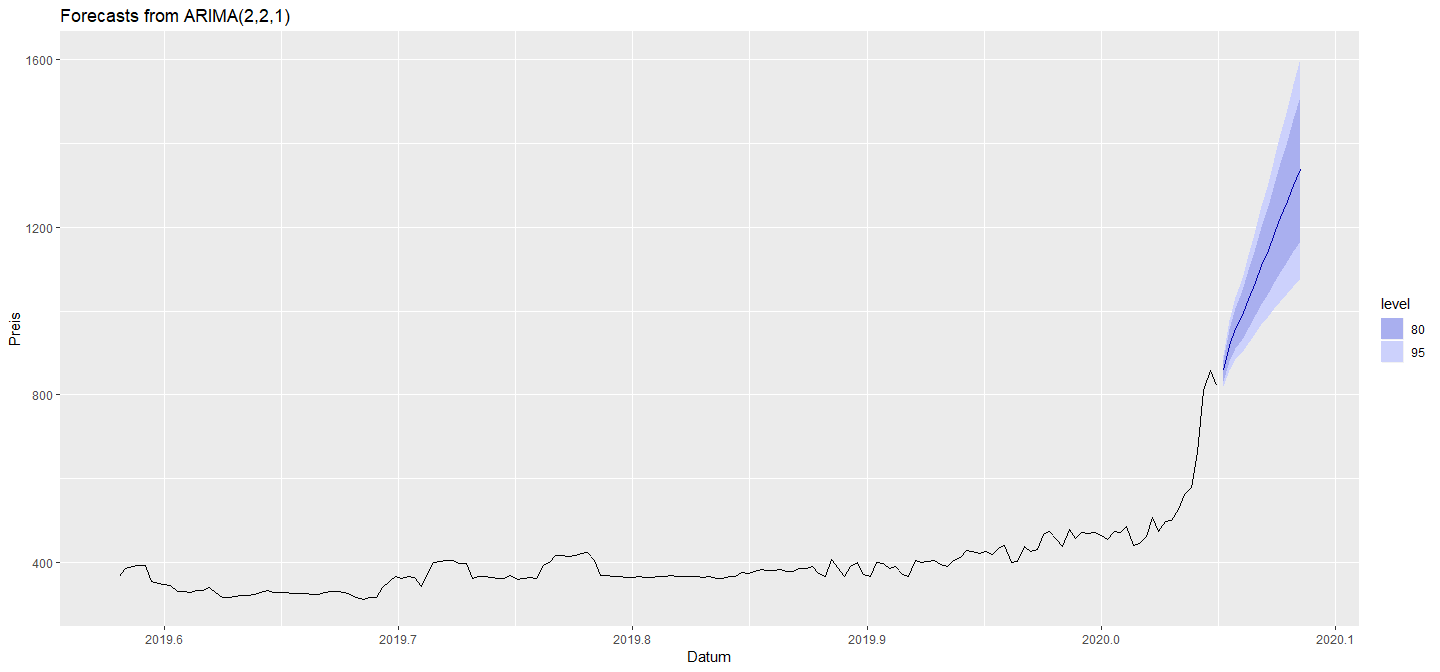
**Diese Zeitreihe lässt sich mit einem ARIMA(2,2,1)-Modell modellieren.**

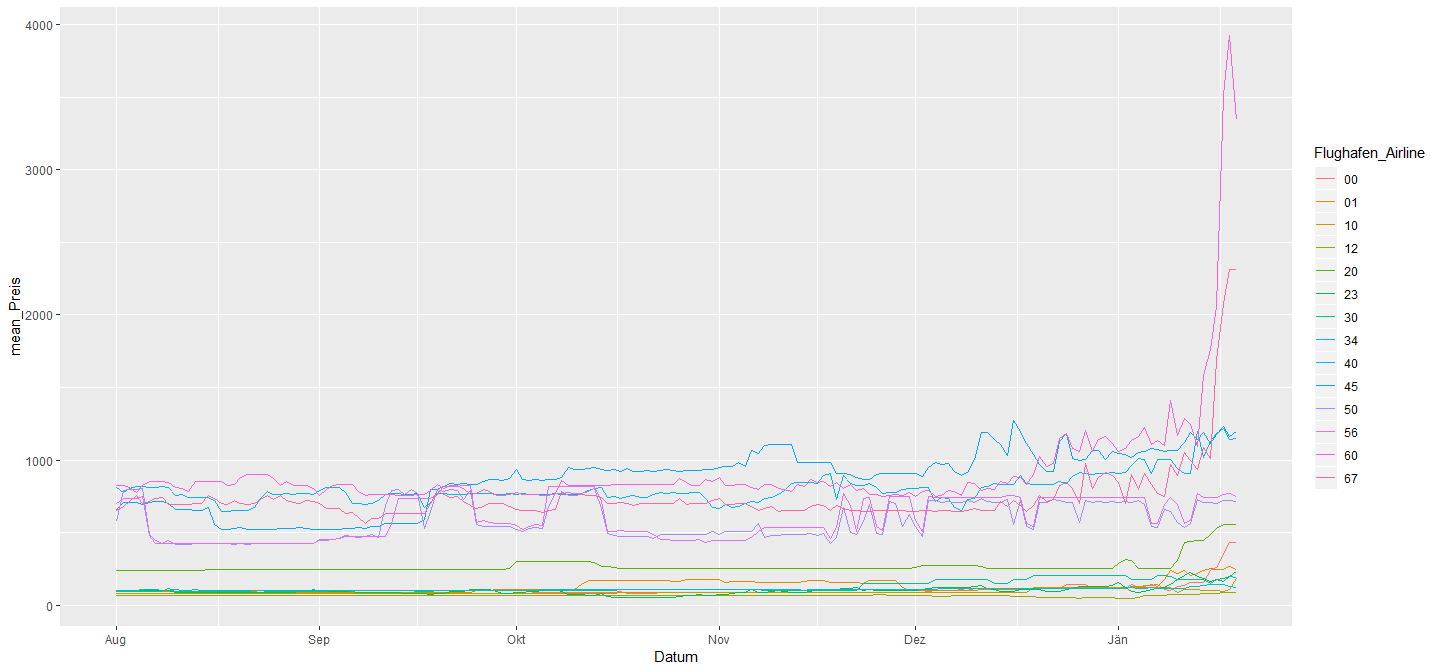
## Series: myts   
## ARIMA(2,2,1)   
##   
## Coefficients:  
## ar1 ar2 ma1  
## 0.0307 -0.2995 -0.7935  
## s.e. 0.1294 0.1130 0.1055  
##   
## sigma^2 estimated as 367.7: log likelihood=-742.56  
## AIC=1493.13 AICc=1493.37 BIC=1505.67  
##   
## Training set error measures:  
## ME RMSE MAE MPE MAPE MASE  
## Training set 1.109718 18.89451 11.46997 0.1191183 2.677086 NaN  
## ACF1  
## Training set 0.000616056

##   
## z test of coefficients:  
##   
## Estimate Std. Error z value Pr(>|z|)   
## ar1 0.03068 0.12939 0.2371 0.812574   
## ar2 -0.29945 0.11300 -2.6501 0.008047 \*\*   
## ma1 -0.79346 0.10548 -7.5222 5.387e-14 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

**Mit diesem Modell kann in weiterer Folge die Zeitrehe angenähert werden und bis zum 1.Februar 2020 vorhergesagt werden.**







### Interpretation - H2

Man kann in allen Plot gut erkennen, dass der Preisverlauf annähernd einem exponentiellen Trend folgt. Dieser Trend (wie im letzten Plot gezeigt) zu einem erheblichen Anteil von den rasanten Preissteigerungen der Flüge 67 und 60 getragen. Das exponentielle Wachstum in der nähe des Abflugsdatums beweist auch unser Forecast, der weiter Preissteigerungen bis zum 1.Februar 2020 vorhersagt. Da die Koeffizienten des ARIMA(2,2,1)-Modells signifkant sind besteht Grund zur Annahme, dass die Preise bei kurzfristigem Buchen stark steigen. Das heißt, die Hypothese 2 wird beibehalten.

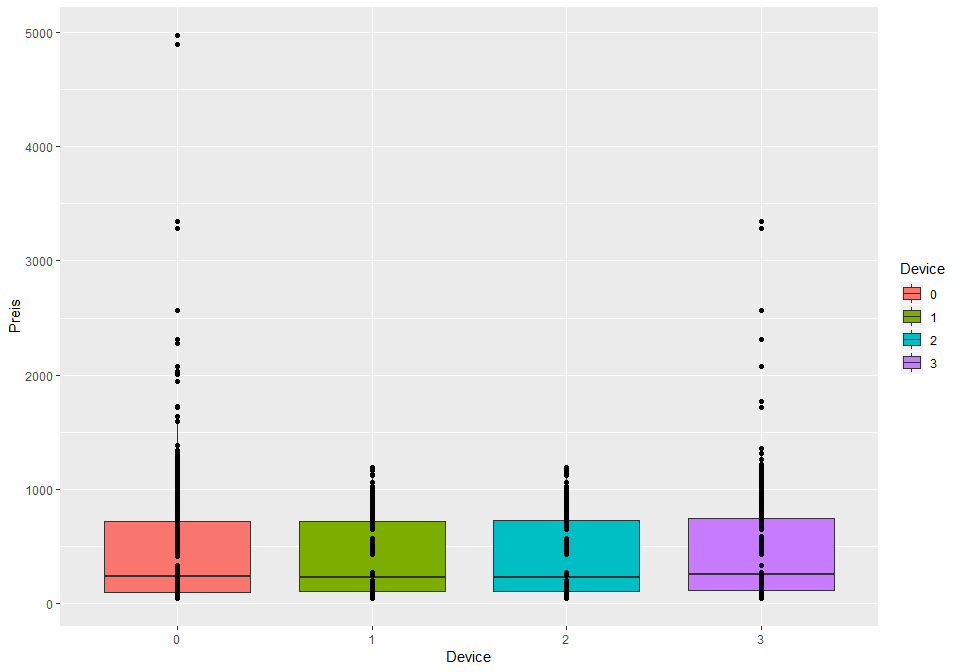
## Hypothese 3

**[Die Wahl des Betriebssystems respektive die Marke des Nutzerendgeräts mit dem die Reise-Website abgerufen wird, hat eine Auswirkung auf den offerierten Preis einer Airline.]**

### Deskreptive Statistik:

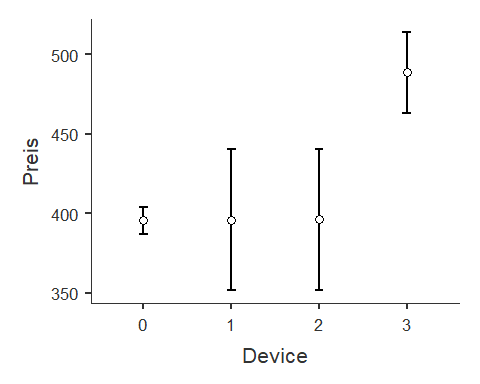
##   
## DESCRIPTIVES  
##   
## Descriptives   
## -----------------------------------------   
## Preis Device   
## -----------------------------------------   
## N 8652 8652   
## Missing 0 0   
## Mean 405   
## Median 248   
## Standard deviation 380   
## Minimum 48.0   
## Maximum 4972   
## -----------------------------------------   
##   
##   
## FREQUENCIES  
##   
## Frequencies of Device   
## --------------------------------------------------   
## Levels Counts % of Total Cumulative %   
## --------------------------------------------------   
## 0 7224 83.5 83.5   
## 1 280 3.2 86.7   
## 2 280 3.2 90.0   
## 3 868 10.0 100.0   
## --------------------------------------------------

### Verteilung - BoxPlot



### Einfaktorielle ANOVA:

##   
## ANOVA  
##   
## ANOVA   
## ------------------------------------------------------------------------   
## Sum of Squares df Mean Square F p   
## ------------------------------------------------------------------------   
## Device 6753467 3 2251156 15.7 < .001   
## Residuals 1.24e+9 8648 143760   
## ------------------------------------------------------------------------   
##   
##   
## ESTIMATED MARGINAL MEANS  
##   
## DEVICE  
##   
## Estimated Marginal Means - Device   
## ---------------------------------------------   
## Device Mean SE Lower Upper   
## ---------------------------------------------   
## 0 395 4.46 387 404   
## 1 396 22.66 351 440   
## 2 396 22.66 352 441   
## 3 489 12.87 463 514   
## ---------------------------------------------



### Interpretation - H3

Nach Emm ist zu beobachten, dass der Mittelwert bei **Macbooks (3)**, ceteris paribus, wesentlich höher ist als bei den anderen Gruppen. Wenn man **FREQUENCIES** betrachtet, dann ist wieder zu erkennen, dass es große Unterschiede in den Sample-Größen der einzelnen Device-Daten gibt. Die ANOVA bestätigt, dass es signifikante Unterschiede zwischen den Gruppen gibt. Das heißt, die Hypothese 3 bestätigt.

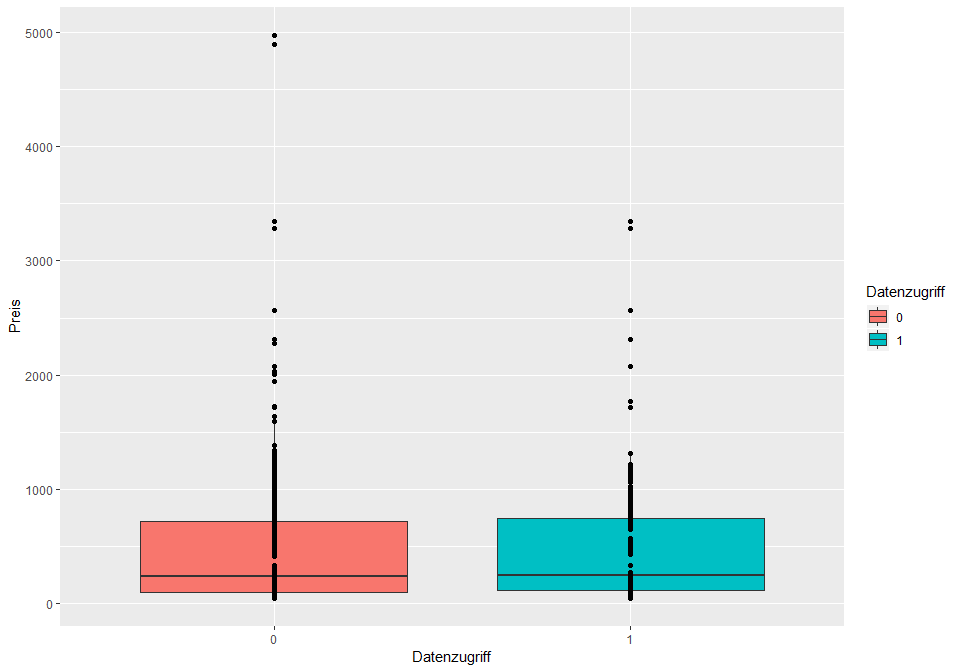
## Hypothese 4

**[Hypothese 4: Das Abrufen einer Reise-Website mittels Applikation und Website erwirkt einen Unterschied des offerierten Preises einer Airline.]**

### Deskreptive Statistik:

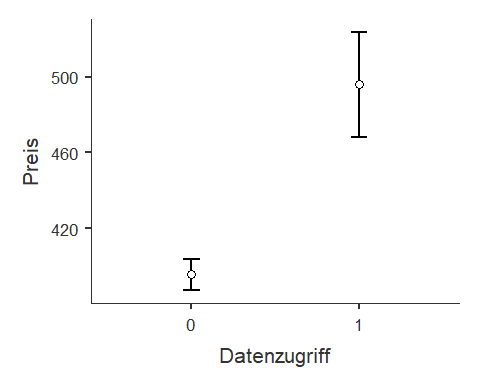
##   
## DESCRIPTIVES  
##   
## Descriptives   
## -----------------------------------------------   
## Preis Datenzugriff   
## -----------------------------------------------   
## N 8512 8512   
## Missing 0 0   
## Mean 404   
## Median 241   
## Standard deviation 380   
## Minimum 48.0   
## Maximum 4972   
## -----------------------------------------------   
##   
##   
## FREQUENCIES  
##   
## Frequencies of Datenzugriff   
## --------------------------------------------------   
## Levels Counts % of Total Cumulative %   
## --------------------------------------------------   
## 0 7784 91.4 91.4   
## 1 728 8.6 100.0   
## --------------------------------------------------

### Verteilung - BoxPlot



### Einfaktorielle ANOVA:

##   
## ANOVA  
##   
## ANOVA   
## ------------------------------------------------------------------------------------   
## Sum of Squares df Mean Square F p <U+03B7>²p   
## ------------------------------------------------------------------------------------   
## Datenzugriff 6709368 1 6709368 46.7 < .001 0.005   
## Residuals 1.22e+9 8510 143817   
## ------------------------------------------------------------------------------------   
##   
##   
## ESTIMATED MARGINAL MEANS  
##   
## DATENZUGRIFF  
##   
## Estimated Marginal Means - Datenzugriff   
## ---------------------------------------------------   
## Datenzugriff Mean SE Lower Upper   
## ---------------------------------------------------   
## 0 396 4.30 387 404   
## 1 496 14.06 468 523   
## ---------------------------------------------------



### Interpretation - H4

In der deskreptiven Statistik haben wir wieder ein große Ungleichheit der Sample-Größen. Allerdings lässt sich ein signifikanter Unterschied im Preis zwischen den zwei Gruppen feststellen. Der Emm-Table zeigt, dass der niedrigste Preis im Schnitt, ceteris paribus, mit **Applications (0)** erzielt wird. Allerdings sei erwähnt, dass sich die Sample-Größen wieder auf den Standardfehler der 2 Gruppen auswirken. Es besteht Beweis für die Hypothese 4. Diese wird daher beibehalten.

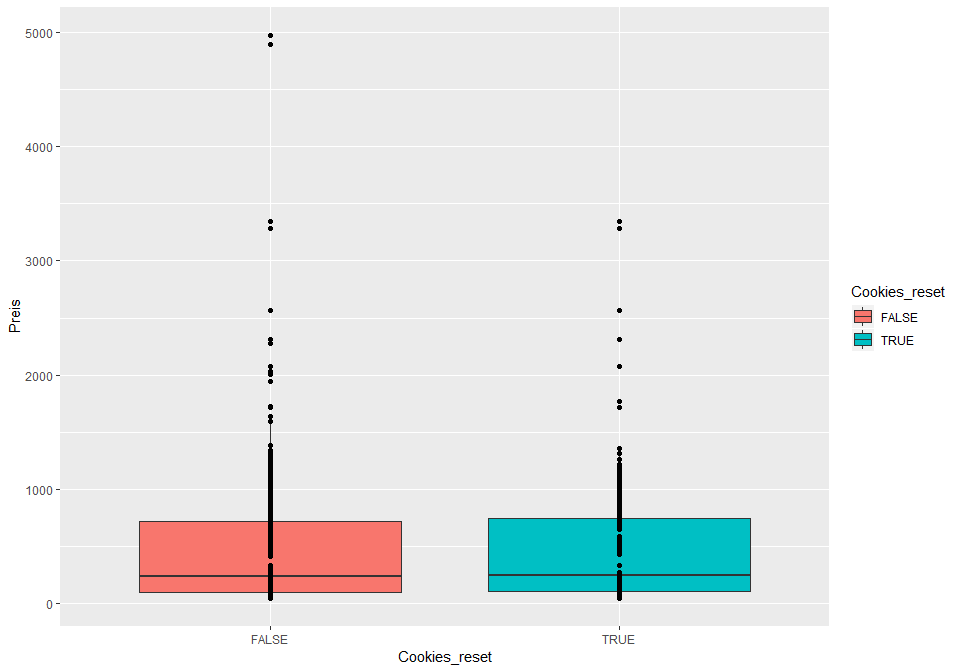
## Hypothese 5

**[Das Zurücksetzen von Cookies respektive dem Browserverlauf erwirkt ein Sinken des offerierten Preises einer Airline.]**

### Deskreptive Statistik:

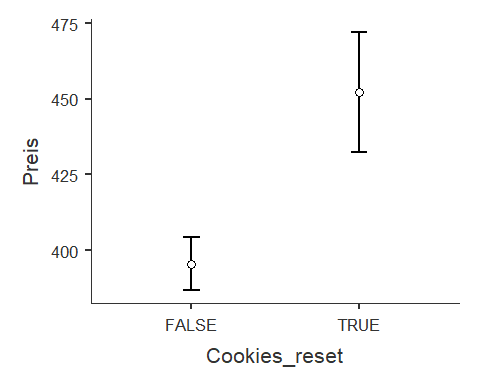
##   
## DESCRIPTIVES  
##   
## Descriptives   
## ------------------------------------------------   
## Preis Cookies\_reset   
## ------------------------------------------------   
## N 8652 8652   
## Missing 0 0   
## Mean 405   
## Median 248   
## Standard deviation 380   
## Minimum 48.0   
## Maximum 4972   
## ------------------------------------------------   
##   
##   
## FREQUENCIES  
##   
## Frequencies of Cookies\_reset   
## --------------------------------------------------   
## Levels Counts % of Total Cumulative %   
## --------------------------------------------------   
## FALSE 7224 83.5 83.5   
## TRUE 1428 16.5 100.0   
## --------------------------------------------------

### Verteilung - BoxPlot



### Einfaktorielle ANOVA:

##   
## ANOVA  
##   
## ANOVA   
## -------------------------------------------------------------------------------------   
## Sum of Squares df Mean Square F p <U+03B7>²p   
## -------------------------------------------------------------------------------------   
## Cookies\_reset 3842588 1 3842588 26.7 < .001 0.003   
## Residuals 1.25e+9 8650 144063   
## -------------------------------------------------------------------------------------   
##   
##   
## ESTIMATED MARGINAL MEANS  
##   
## COOKIES\_RESET  
##   
## Estimated Marginal Means - Cookies\_reset   
## ----------------------------------------------------   
## Cookies\_reset Mean SE Lower Upper   
## ----------------------------------------------------   
## FALSE 395 4.47 387 404   
## TRUE 452 10.04 433 472   
## ----------------------------------------------------



### Interpretation - H5

Zunächst gibt es wieder eine große Ungleichheit zwischen den erfassten Cookie-Daten. Zu beachten ist, dass die Variable cookies\_reset eine binäre Variable darstellt und die Codierung TRUE = ‘zurückgesetzt’ und FALSE = ‘Zugelassen’ beinhaltet. Nach der einfaktoriellen ANOVA ist nach dem p-Wert die Nullhypothese zu verwerfen. Das heißt, es gibt einen signifikanten Unterschied zwischen den zwei Gruppen, ceteris paribus. Bei Betrachtung des Emm ist ersichtlich, dass das **Zurücksetzen** der Cookies mit einem höheren mittleren Preis verbunden ist.

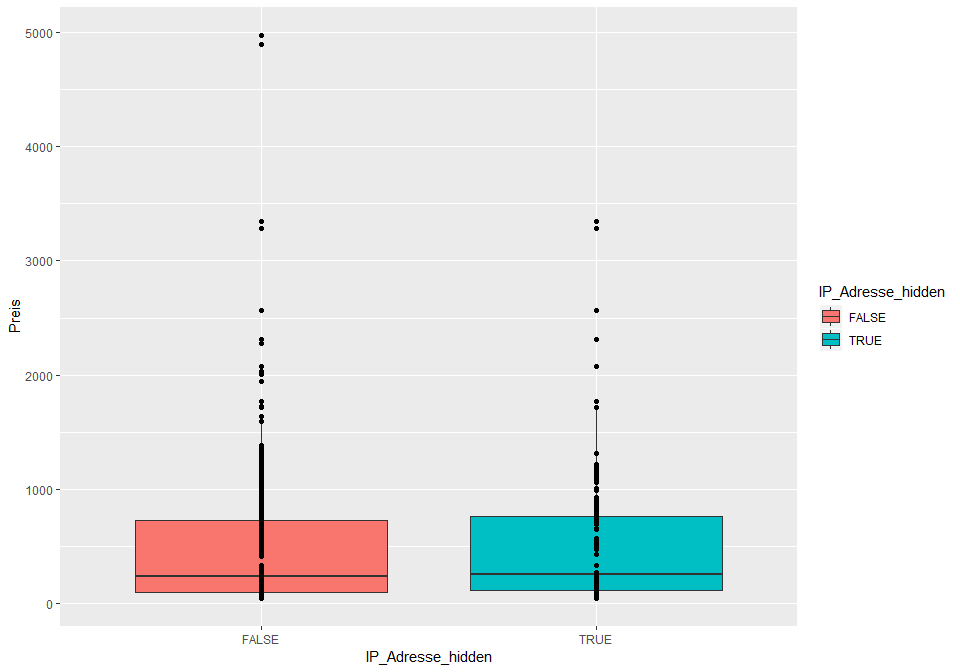
## Hypothese 6

**[Das Verbergen der Internetprotokoll-Adresse und folglich der ortsspezifischen Parameter mittels Virtual Private Network verursacht eine Differenz im offerierten Preis einer Airline.]**

### Deskreptive Statistik:

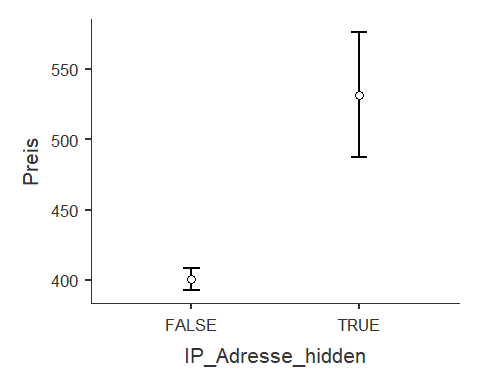
##   
## DESCRIPTIVES  
##   
## Descriptives   
## ----------------------------------------------------   
## Preis IP\_Adresse\_hidden   
## ----------------------------------------------------   
## N 8652 8652   
## Missing 0 0   
## Mean 405   
## Median 248   
## Standard deviation 380   
## Minimum 48.0   
## Maximum 4972   
## ----------------------------------------------------   
##   
##   
## FREQUENCIES  
##   
## Frequencies of IP\_Adresse\_hidden   
## --------------------------------------------------   
## Levels Counts % of Total Cumulative %   
## --------------------------------------------------   
## FALSE 8372 96.8 96.8   
## TRUE 280 3.2 100.0   
## --------------------------------------------------

### Verteilung - BoxPlot



### Einfaktorielle ANOVA:

##   
## ANOVA  
##   
## ANOVA   
## -----------------------------------------------------------------------------------------   
## Sum of Squares df Mean Square F p <U+03B7>²p   
## -----------------------------------------------------------------------------------------   
## IP\_Adresse\_hidden 4644280 1 4644280 32.3 < .001 0.004   
## Residuals 1.25e+9 8650 143971   
## -----------------------------------------------------------------------------------------   
##   
##   
## ESTIMATED MARGINAL MEANS  
##   
## IP\_ADRESSE\_HIDDEN  
##   
## Estimated Marginal Means - IP\_Adresse\_hidden   
## --------------------------------------------------------   
## IP\_Adresse\_hidden Mean SE Lower Upper   
## --------------------------------------------------------   
## FALSE 401 4.15 392 409   
## TRUE 532 22.68 487 576   
## --------------------------------------------------------



### Interpretation - H6

Es gibt wieder eine sehr große Ungleichheit zwischen den erfassten IP-Adressen-Daten. Zu beachten ist, dass die Variable IP\_Adresse\_hidden eine binäre Variable darstellt und die Codierung TRUE = ‘verborgen’ und FALSE = ‘sichtbar’ beinhaltet. Nach der einfaktoriellen ANOVA ist nach dem p-Wert die Nullhypothese zu verwerfen. Das heißt, es gibt einen signifikanten Unterschied im Mittelwert der zwei Gruppen, ceteris paribus. Bei Betrachtung des Emm ist ersichtlich, dass das **Verbergen** der Cookies mit einem höheren mittleren Preis verbunden ist.

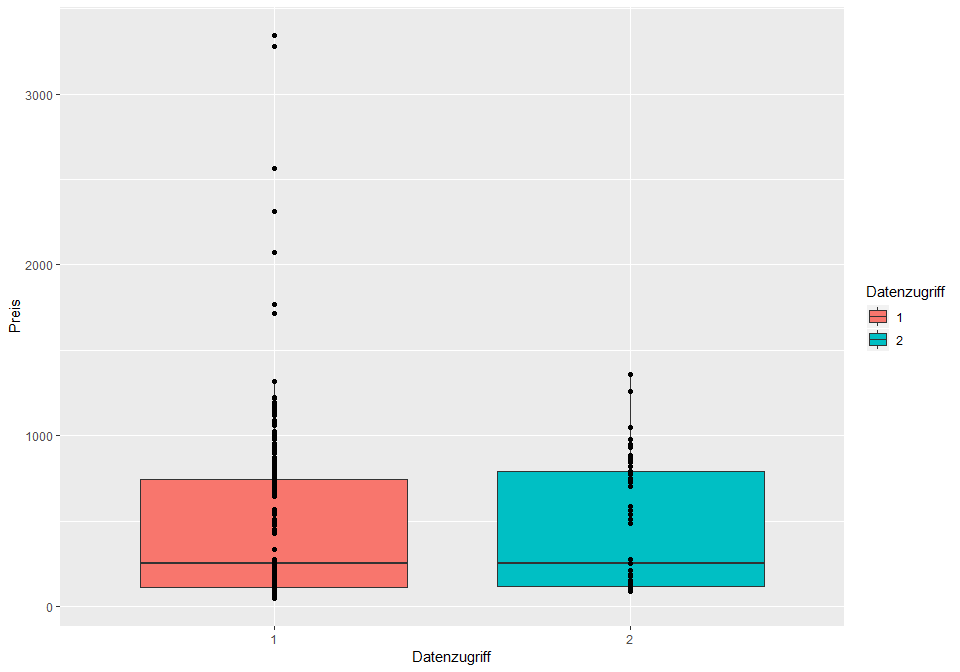
## Hypothese 7

**[Das Abrufen eines Flugpreises via Reise-Website führt, verglichen mit der Website der Airline selbst, zu einem höheren offerierten Preis.]**

### Deskreptive Statistik:

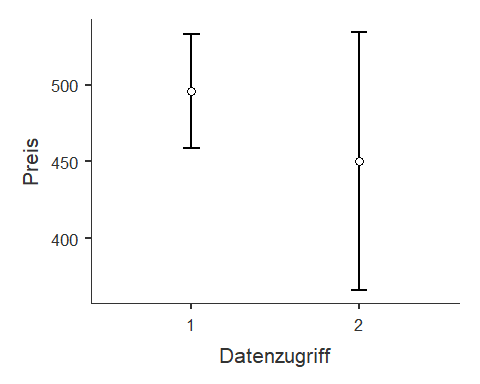
##   
## DESCRIPTIVES  
##   
## Descriptives   
## -----------------------------------------------   
## Preis Datenzugriff   
## -----------------------------------------------   
## N 868 868   
## Missing 0 0   
## Mean 489   
## Median 253   
## Standard deviation 507   
## Minimum 48.0   
## Maximum 3343   
## -----------------------------------------------   
##   
##   
## FREQUENCIES  
##   
## Frequencies of Datenzugriff   
## --------------------------------------------------   
## Levels Counts % of Total Cumulative %   
## --------------------------------------------------   
## 1 728 83.9 83.9   
## 2 140 16.1 100.0   
## --------------------------------------------------

### Verteilung - BoxPlot



### Einfaktorielle ANOVA:

##   
## ANOVA  
##   
## ANOVA   
## -----------------------------------------------------------------------------------   
## Sum of Squares df Mean Square F p <U+03B7>²p   
## -----------------------------------------------------------------------------------   
## Datenzugriff 246886 1 246886 0.960 0.327 0.001   
## Residuals 2.23e+8 866 257166   
## -----------------------------------------------------------------------------------   
##   
##   
## ESTIMATED MARGINAL MEANS  
##   
## DATENZUGRIFF  
##   
## Estimated Marginal Means - Datenzugriff   
## --------------------------------------------------   
## Datenzugriff Mean SE Lower Upper   
## --------------------------------------------------   
## 1 496 18.8 459 533   
## 2 450 42.9 366 534   
## --------------------------------------------------



### Interpretation - H7

Die Unterschiede sind nicht signfikant. Das heißt, die Nullhypothese wird beibehalten. Es konnte nicht nachgewiesen werden, dass es unterschiedliche Preise für **Website (1)** und **Reise-Website (2)** gibt. Zudem gibt es wieder krasse Disbalancen zwischen den Sample-Größen der Merkmale.