

Silas_Analysis

Max Kuttner

25 3 2020

Hypothese 1

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

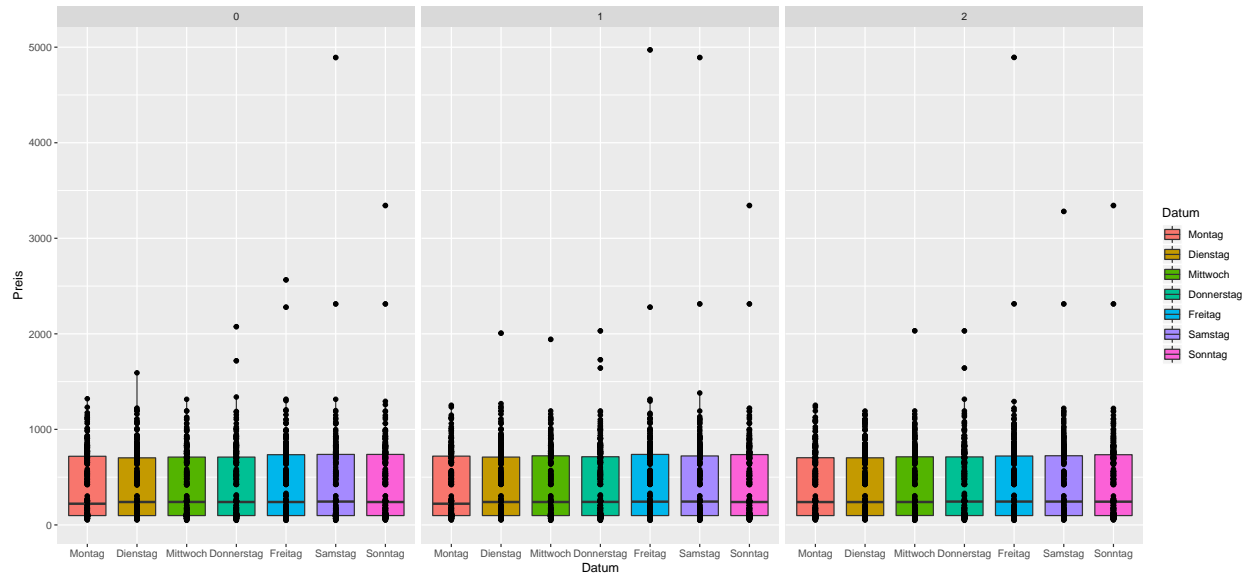
```
## [1] "Montag"      "Dienstag"    "Mittwoch"    "Donnerstag"  "Freitag"
## [6] "Samstag"     "Sonntag"
```

Deskreptive Statistik:

```
##
## DESCRIPTIVES
##
## Descriptives
## -----
##               Preis      Datum      Zeit
## -----
##      N              7224      7224      7224
##      Missing          0          0          0
##      Mean             395
##      Median           241
##      Standard deviation 364
##      Minimum          48.0
##      Maximum          4972
## -----
##
##
## FREQUENCIES
##
## Frequencies of Datum
## -----
##      Levels      Counts      % of Total      Cumulative %
## -----
##      Montag        1008          14.0          14.0
##      Dienstag      1008          14.0          27.9
##      Mittwoch       1008          14.0          41.9
##      Donnerstag     1050          14.5          56.4
##      Freitag        1050          14.5          70.9
##      Samstag        1050          14.5          85.5
##      Sonntag        1050          14.5          100.0
## -----
##
##
## Frequencies of Zeit
## -----
##      Levels      Counts      % of Total      Cumulative %
## -----
##      0            2408          33.3          33.3
```

```
##      1      2408      33.3      66.7
##      2      2408      33.3     100.0
## -----
```

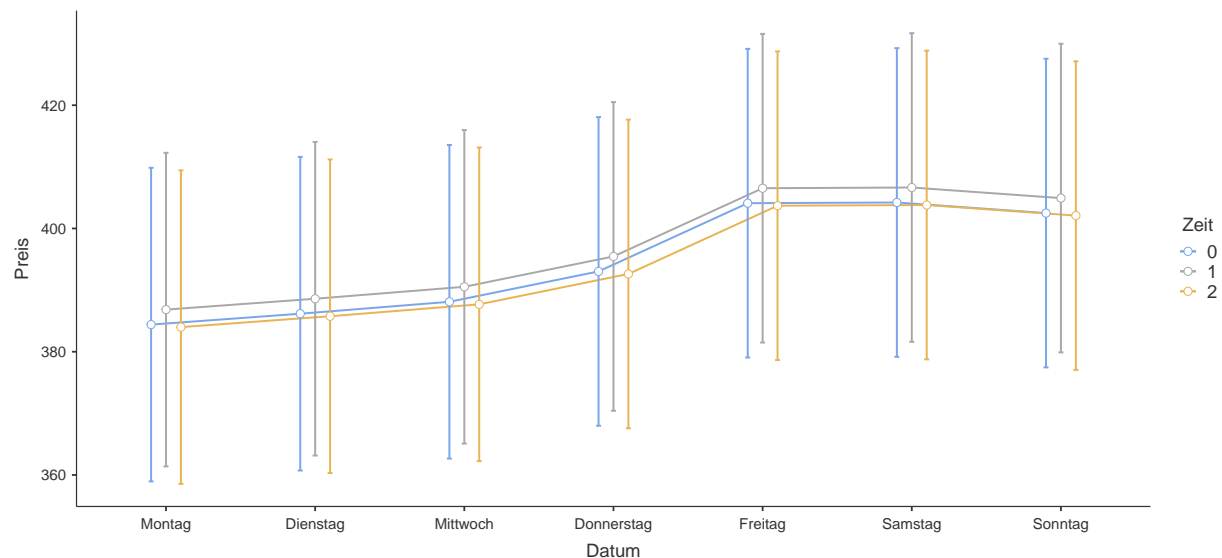
Verteilung - BoxPlot



Mehrfaktorielle ANOVA:

```
##
## ANOVA
##
## ANOVA
## -----
##              Sum of Squares    df    Mean Square    F        p        <U+03B7>²p
## -----
## Datum              478813         6         79802    0.6011    0.730    0.000
## Zeit               11357          2         5678     0.0428    0.958    0.000
## Residuals         9.58e+8       7215        132767
## -----
##
## ESTIMATED MARGINAL MEANS
##
## DATUM:ZEIT
##
## Estimated Marginal Means - Datum:Zeit
## -----
##      Zeit    Datum      Mean    SE    Lower    Upper
## -----
##      0      Montag      384    13.0    359     410
##      0      Dienstag    386    13.0    361     412
##      0      Mittwoch     388    13.0    363     414
##      0      Donnerstag    393    12.8    368     418
##      0      Freitag      404    12.8    379     429
##      0      Samstag      404    12.8    379     429
```

```
##          Sonntag      403    12.8    377    428
##    1 Montag        387    13.0    361    412
##      Dienstag     389    13.0    363    414
##      Mittwoch      391    13.0    365    416
##      Donnerstag     395    12.8    370    421
##      Freitag       407    12.8    381    432
##      Samstag       407    12.8    382    432
##    2 Sonntag       405    12.8    380    430
##      Montag        384    13.0    359    409
##      Dienstag     386    13.0    360    411
##      Mittwoch      388    13.0    362    413
##      Donnerstag     393    12.8    368    418
##      Freitag       404    12.8    379    429
##      Samstag       404    12.8    379    429
##      Sonntag       402    12.8    377    427
## -----
```



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    384.    13.0  359.  409.
## 2 0 Montag    384.    13.0  359.  410.
## 3 2 Dienstag 386.    13.0  360.  411.
## 4 0 Dienstag 386.    13.0  361.  412.
## 5 1 Montag    387.    13.0  361.  412.
## 6 2 Mittwoch 388.    13.0  362.  413.
## 7 0 Mittwoch 388.    13.0  363.  414.
## 8 1 Dienstag 389.    13.0  363.  414.
## 9 1 Mittwoch 391.    13.0  365.  416.
```

```
## 10 2    Donnerstag 393. 12.8 368. 418.
## # ... with 11 more rows
```

Interpretation - H1

Wir sehen durch die ANOVA, dass beide Gruppen keinen signifikante Preisunterschiede aufweisen. Das heißt die Nullhypothese wird in diesem Fall beibehalten.

Es 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

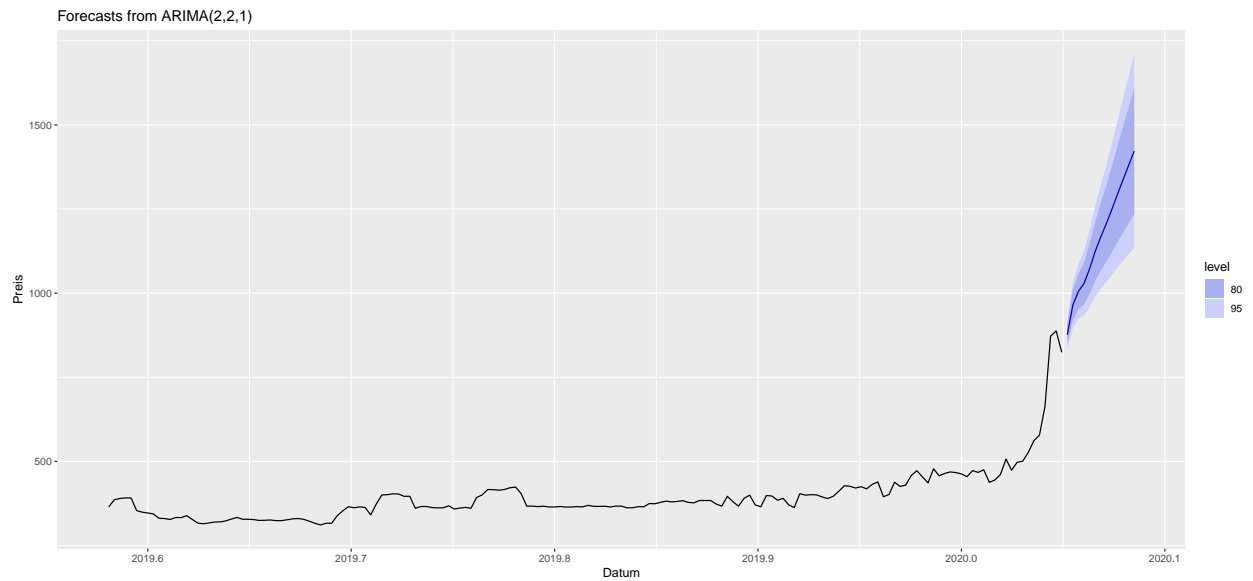
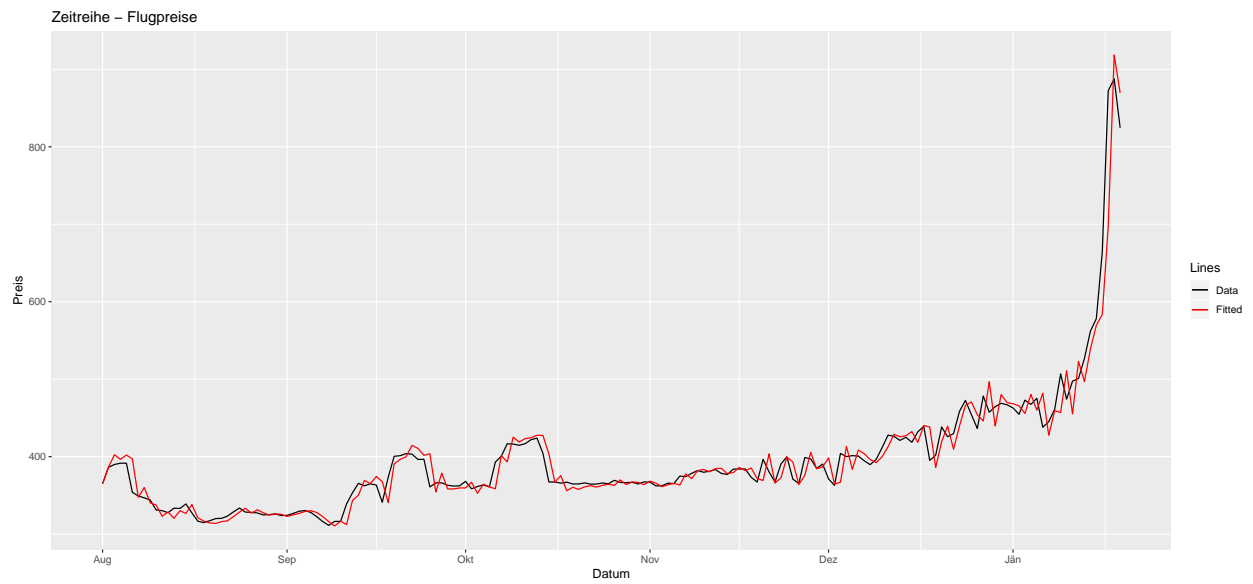
[Je spontaner und kurzfristiger die Kaufentscheidung getroffen wird, desto höher ist der offerierte Preis einer Airline.]

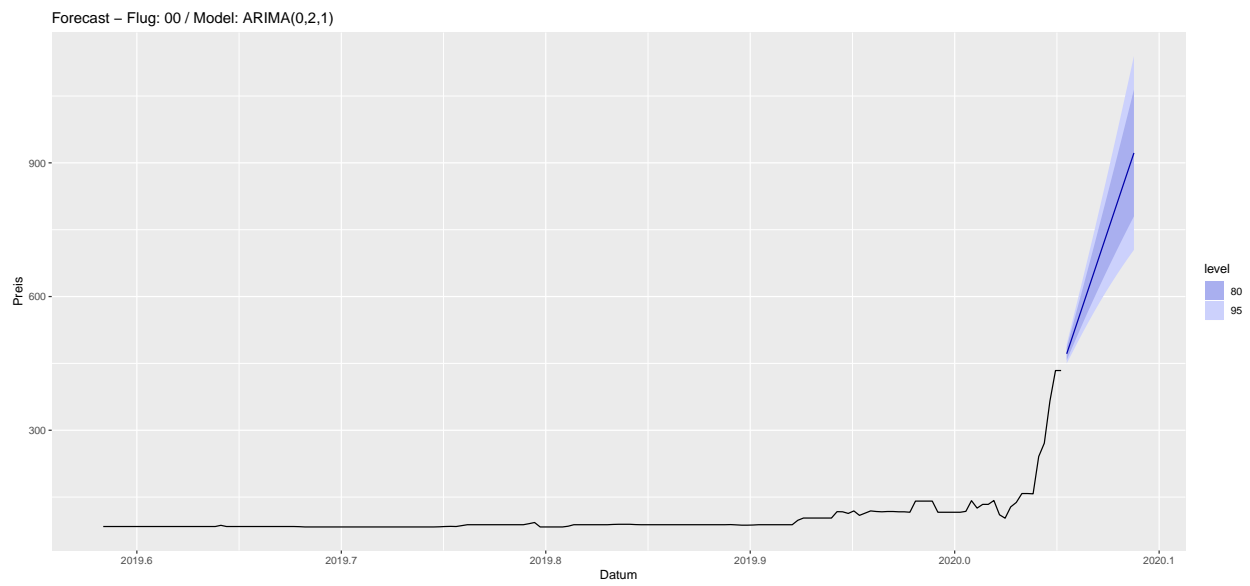
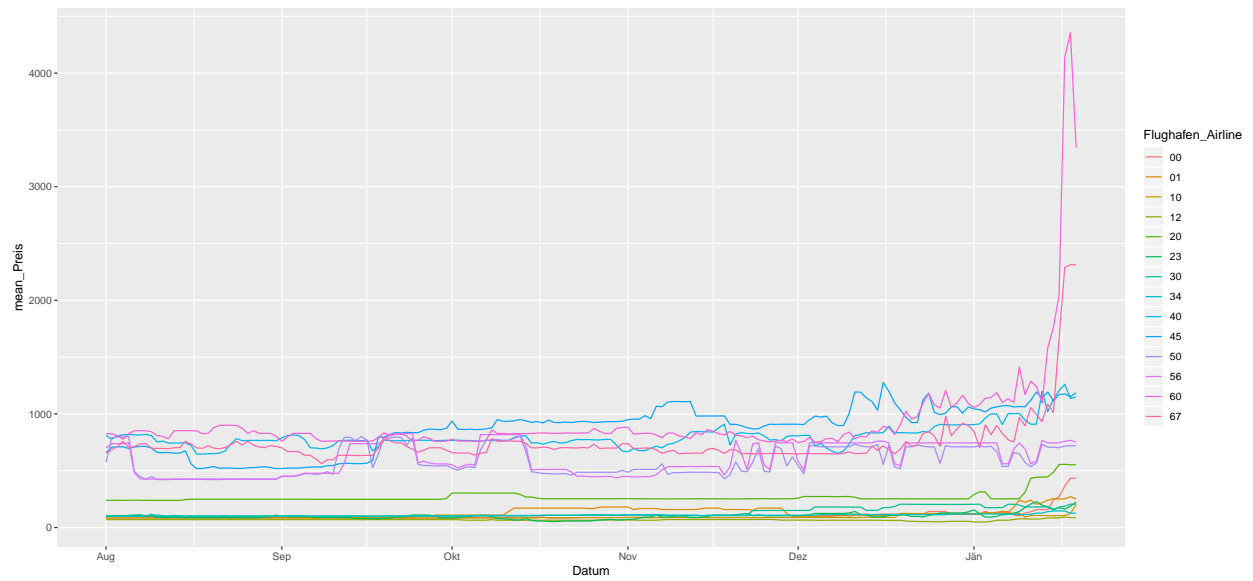
```
## Warning in value[[3L]](cond): The chosen test encountered an error, so no
## seasonal differencing is selected. Check the time series data.
```

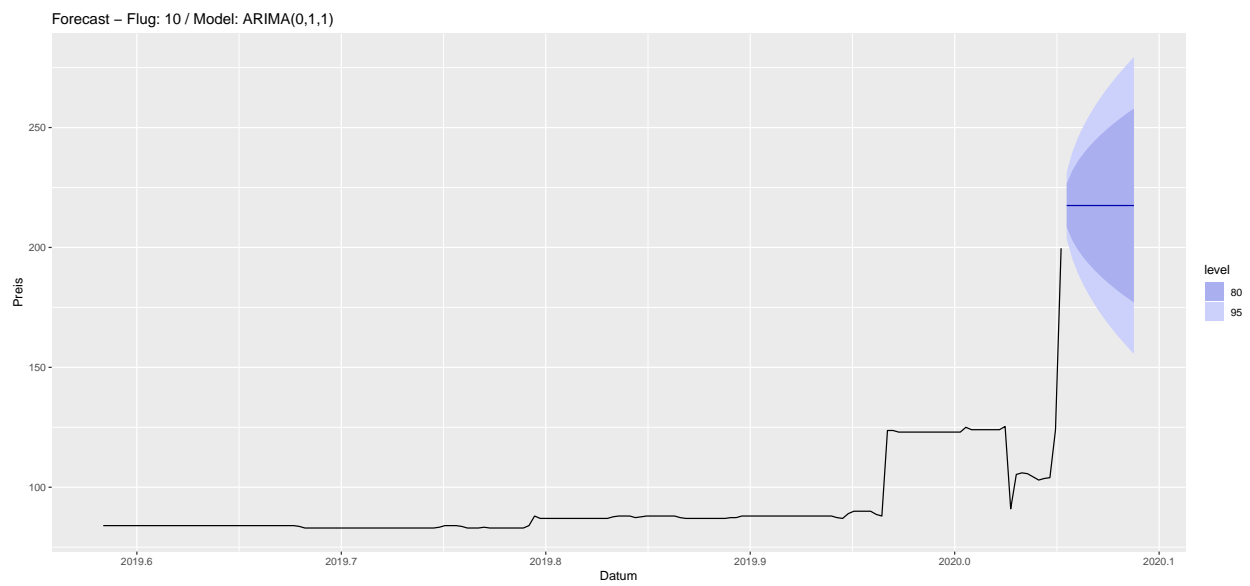
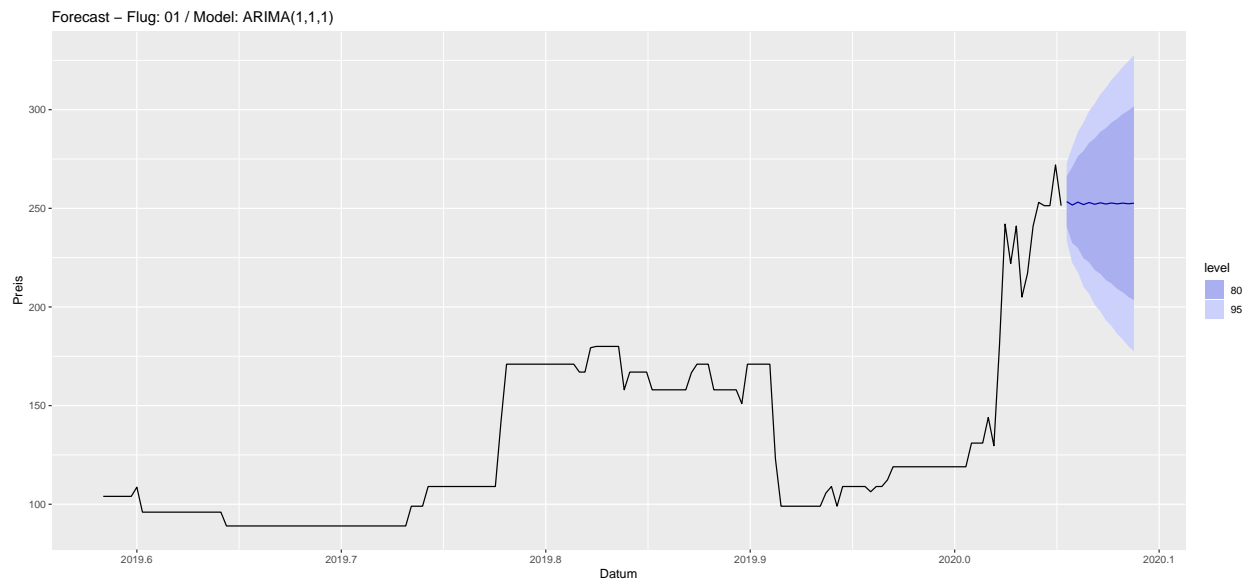
Diese Zeitreihe lässt sich mit einem ARIMA(2,2,1)-Modell modellieren.

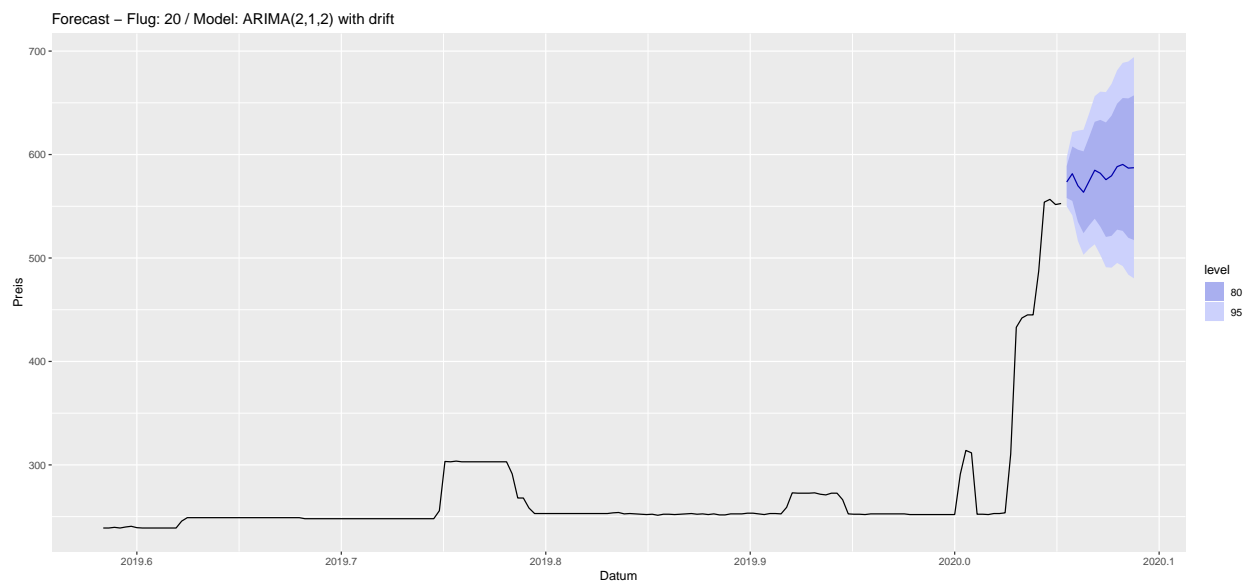
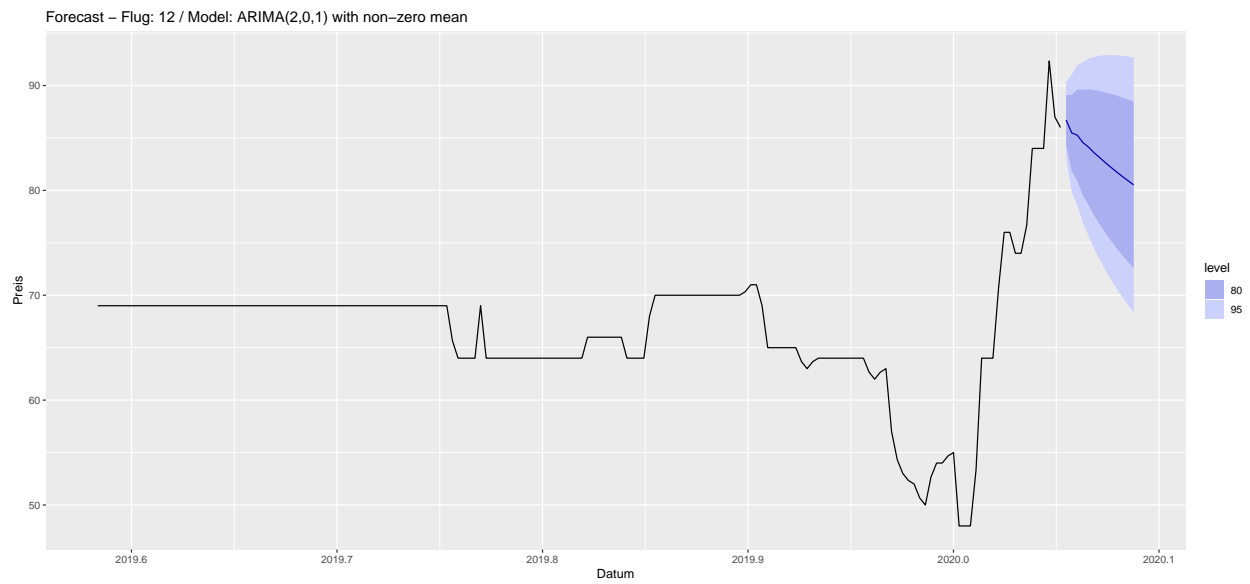
```
## Series: myts
## ARIMA(2,2,1)
##
## Coefficients:
##          ar1      ar2      ma1
##          0.0271  -0.4283  -0.7750
## s.e.  0.1274   0.1189   0.1074
##
## sigma^2 estimated as 467.7:  log likelihood=-763.15
## AIC=1534.3   AICc=1534.54   BIC=1546.84
##
## Training set error measures:
##              ME      RMSE      MAE      MPE      MAPE  MASE      ACF1
## Training set 1.289643 21.31025 11.96192 0.140989 2.742449  NaN  0.005852603
##
## z test of coefficients:
##
##      Estimate Std. Error z value Pr(>|z|)
## ar1  0.02708    0.12737  0.2126 0.8316330
## ar2 -0.42829    0.11888 -3.6027 0.0003149 ***
## ma1 -0.77498    0.10736 -7.2182 5.268e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

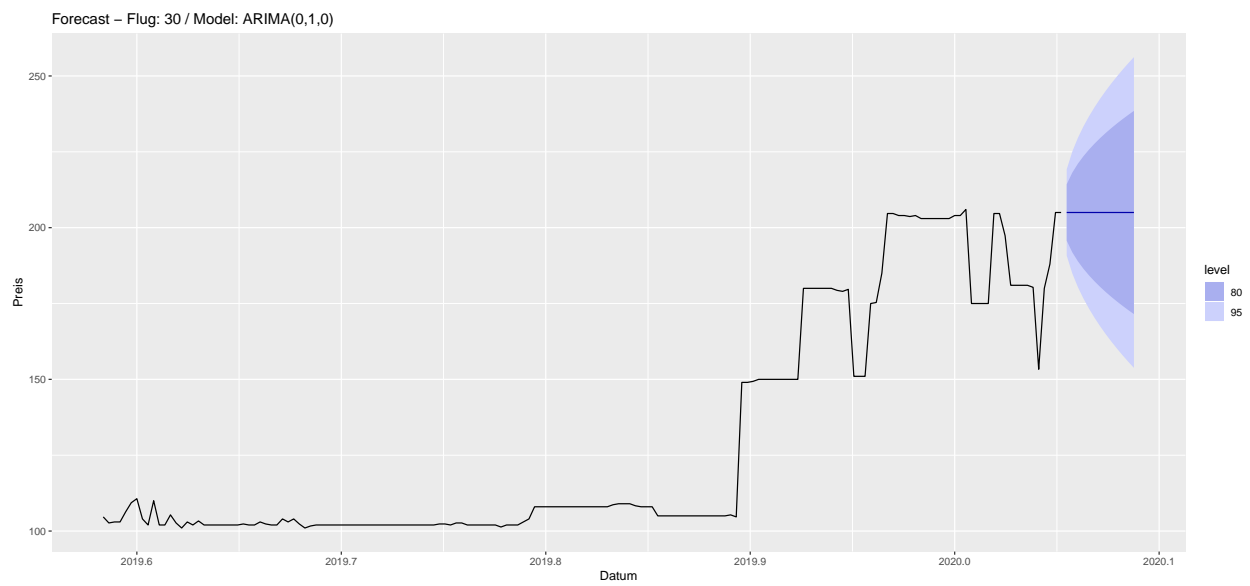
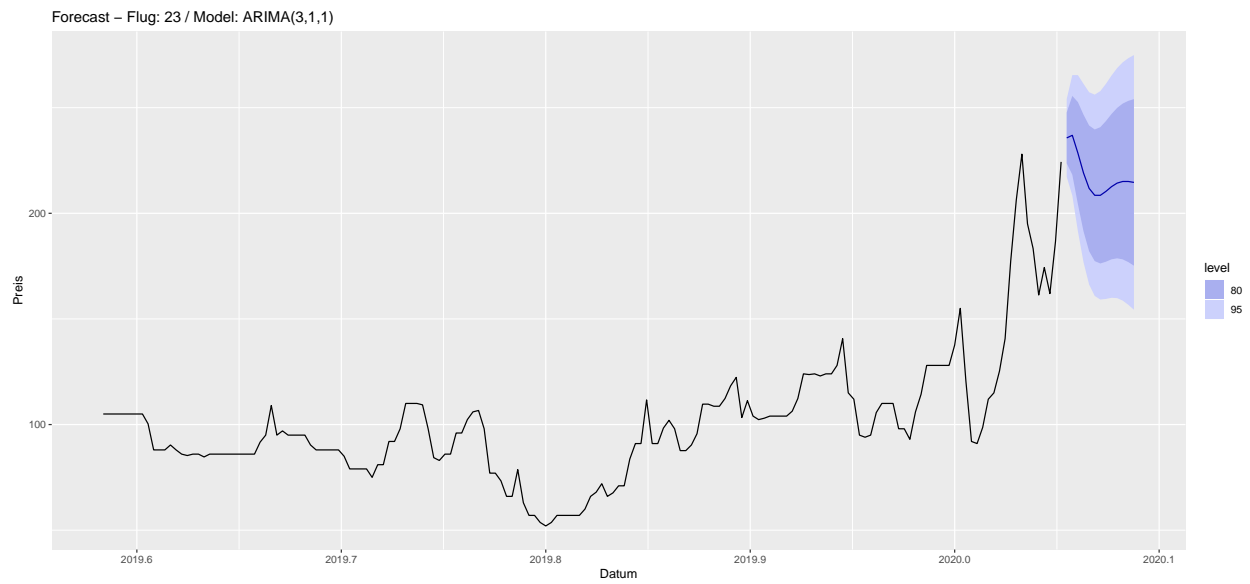
Mit diesem Modell kann in weiterer Folge die Zeitreihe angenähert werden und bis zum 1. Februar 2020 vorhergesagt werden.

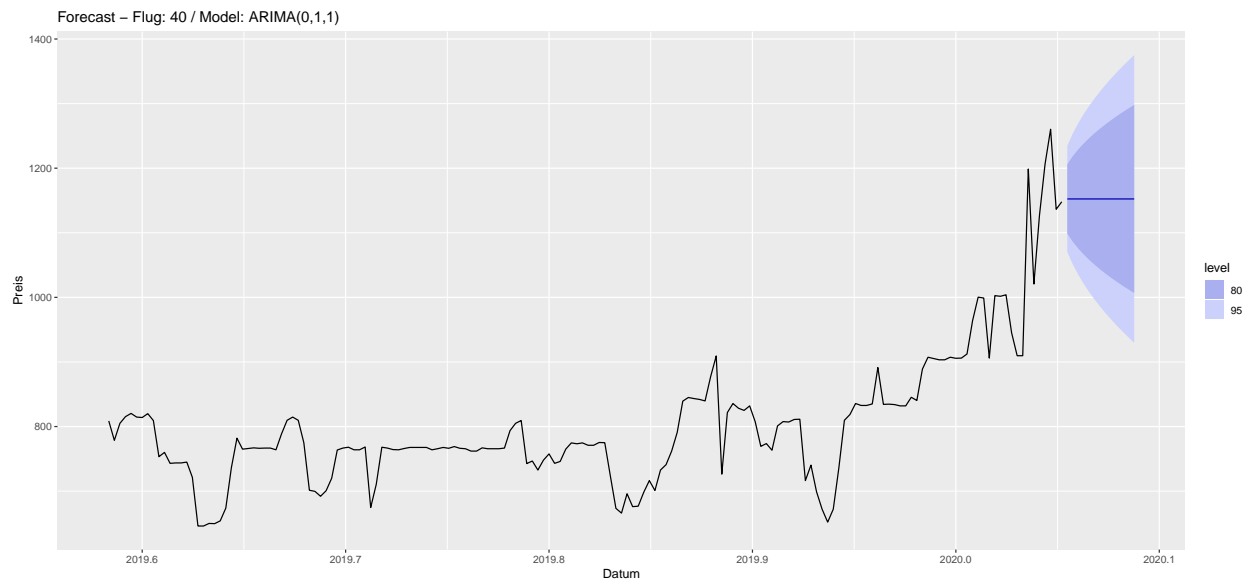
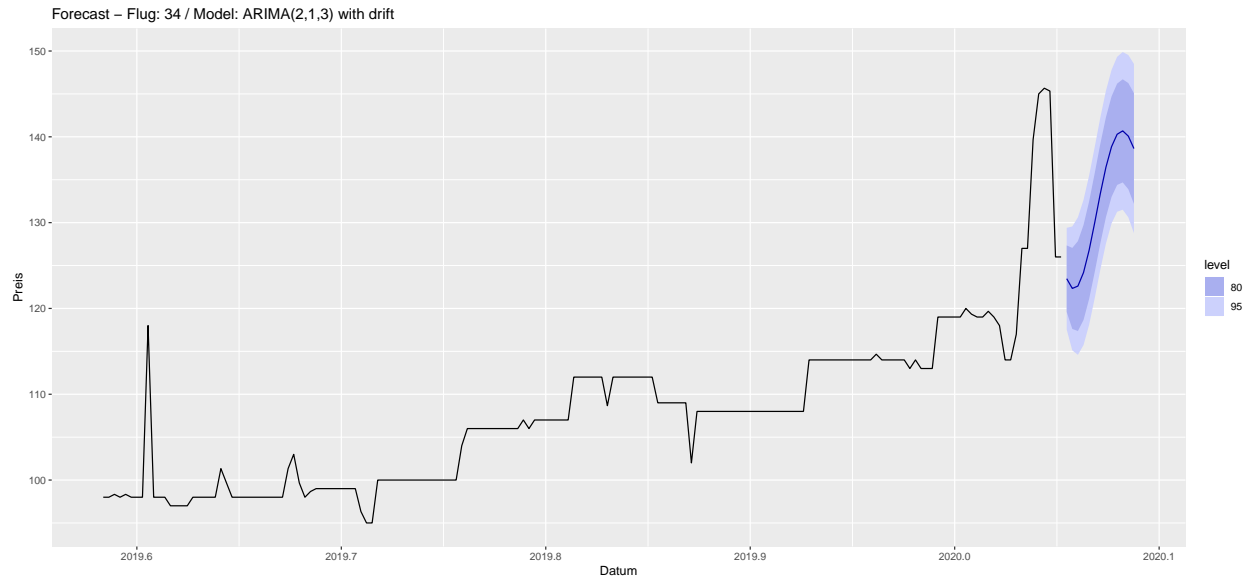


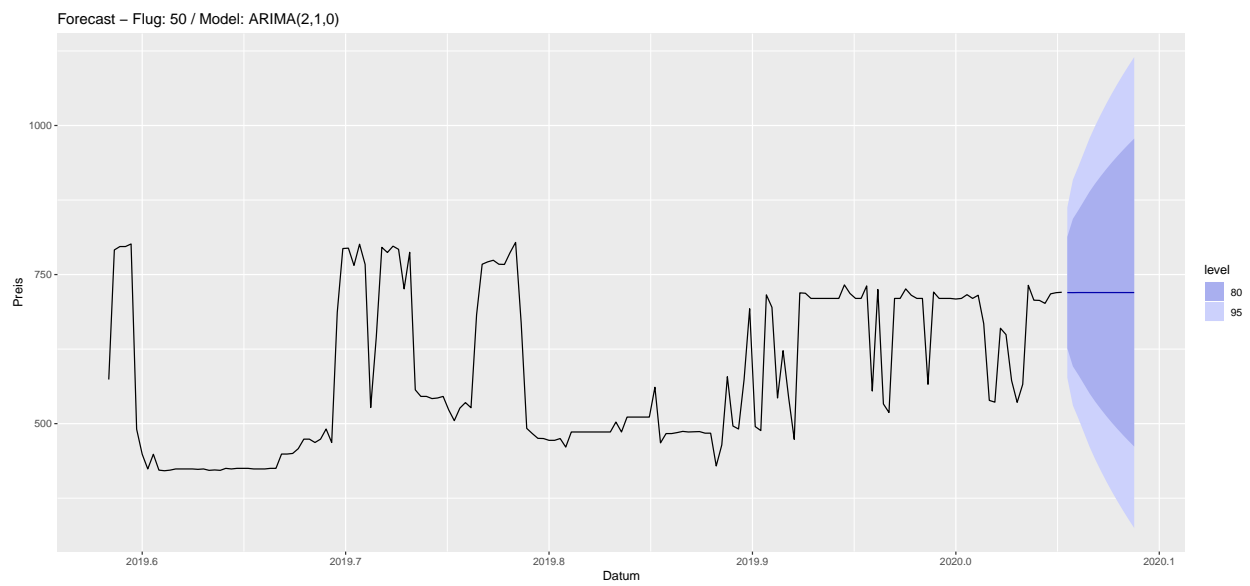
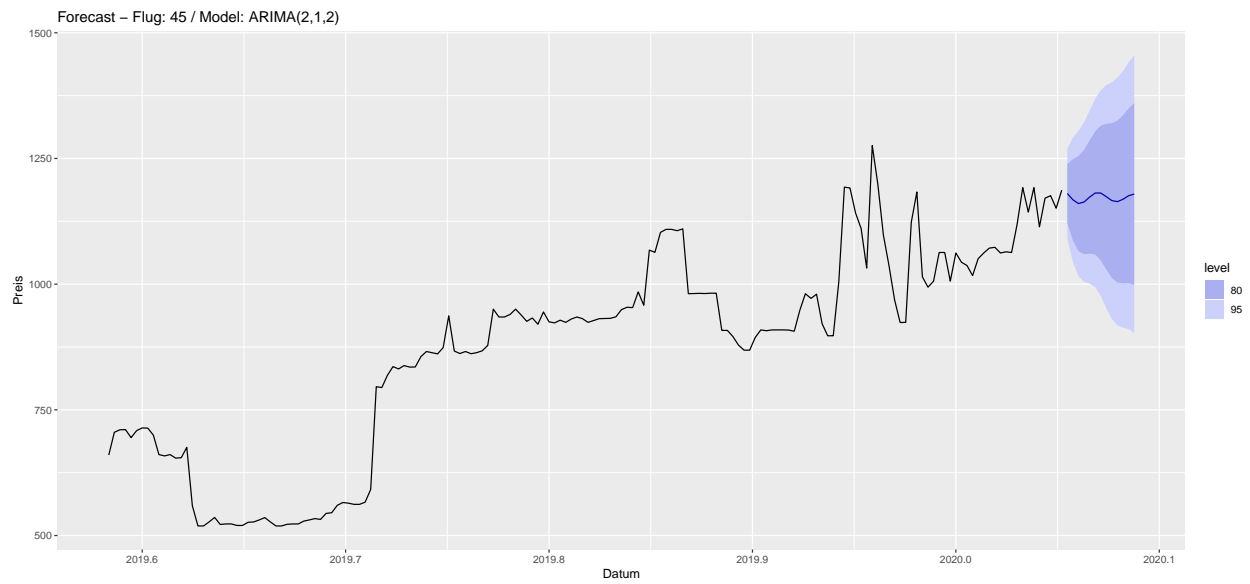


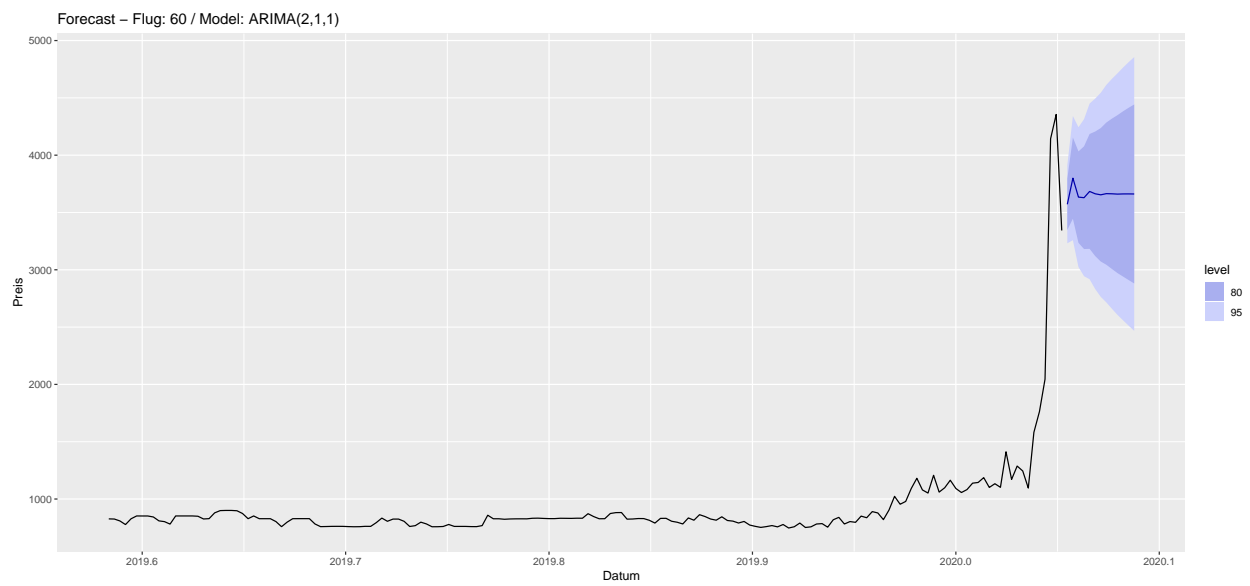
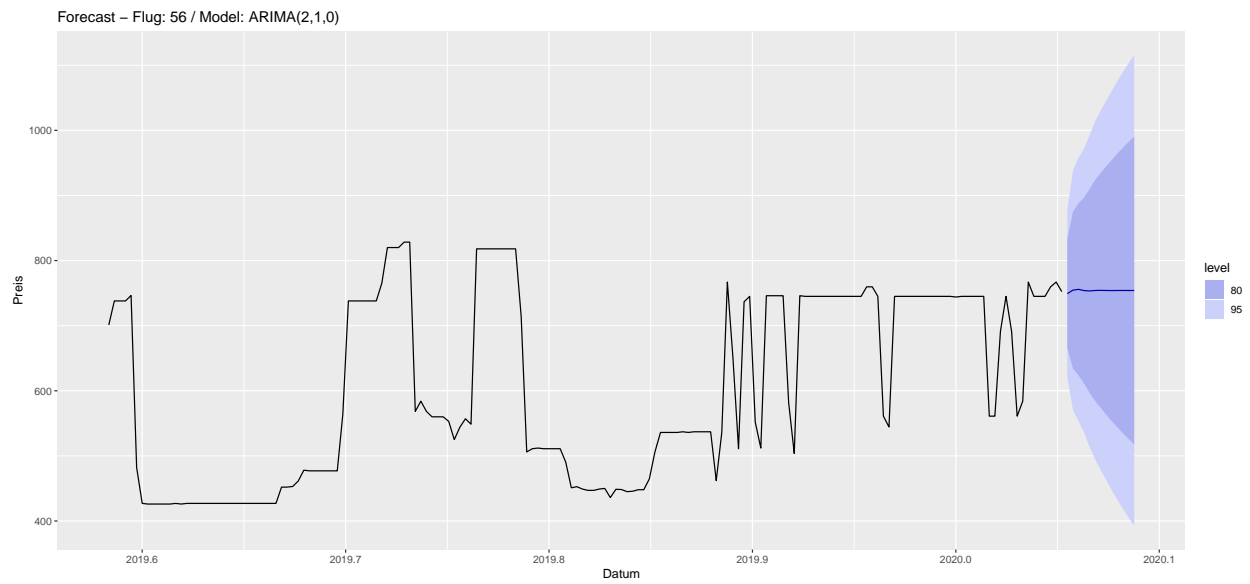


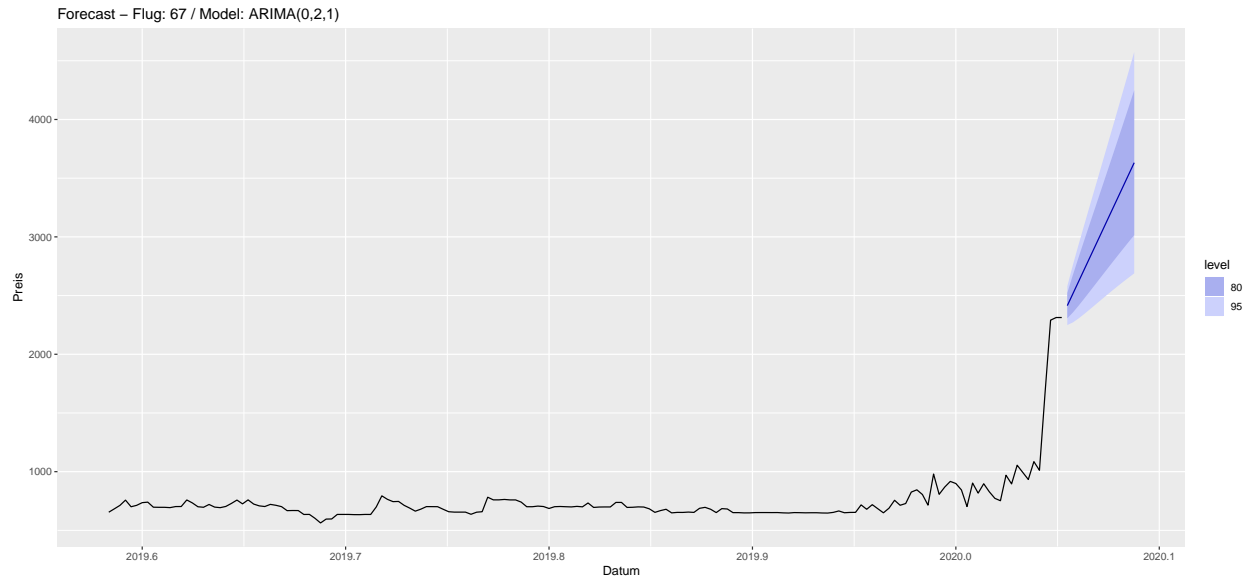












Interpretation - H2

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 signifikant sind besteht Grund zur Annahme, dass die Preise bei kurzfristigem Buchen stark steigen. Das heißt, die Hypothese 2 wird beibehalten.

Hypothese 3

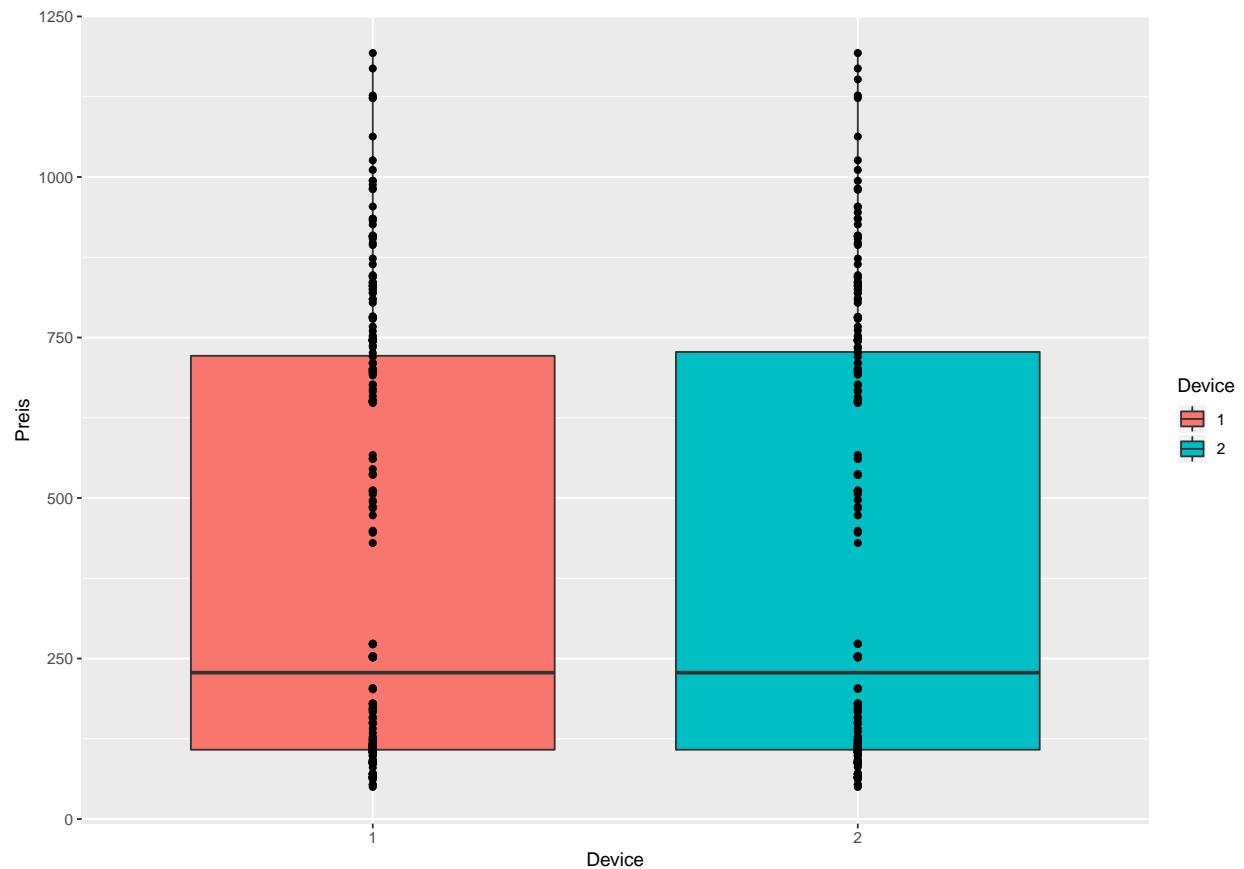
[Die Wahl des Betriebssystems respektive die Marke des Nutzerendgeräts mit dem die Reise-Website abgerufen wird, hat eine Auswirkung auf den offerierten Preis einer Airline.]

Deskreptive Statistik:

```
##
## DESCRIPTIVES
##
## Descriptives
## -----
##               Preis      Device
## -----
## N                560        560
## Missing           0          0
## Mean             396
## Median           228
## Standard deviation 333
## Minimum          50.0
## Maximum          1193
## -----
##
##
## FREQUENCIES
##
```

```
## Frequencies of Device
## -----
##      Levels      Counts      % of Total      Cumulative %
## -----
##      1           280         50.0           50.0
##      2           280         50.0          100.0
## -----
```

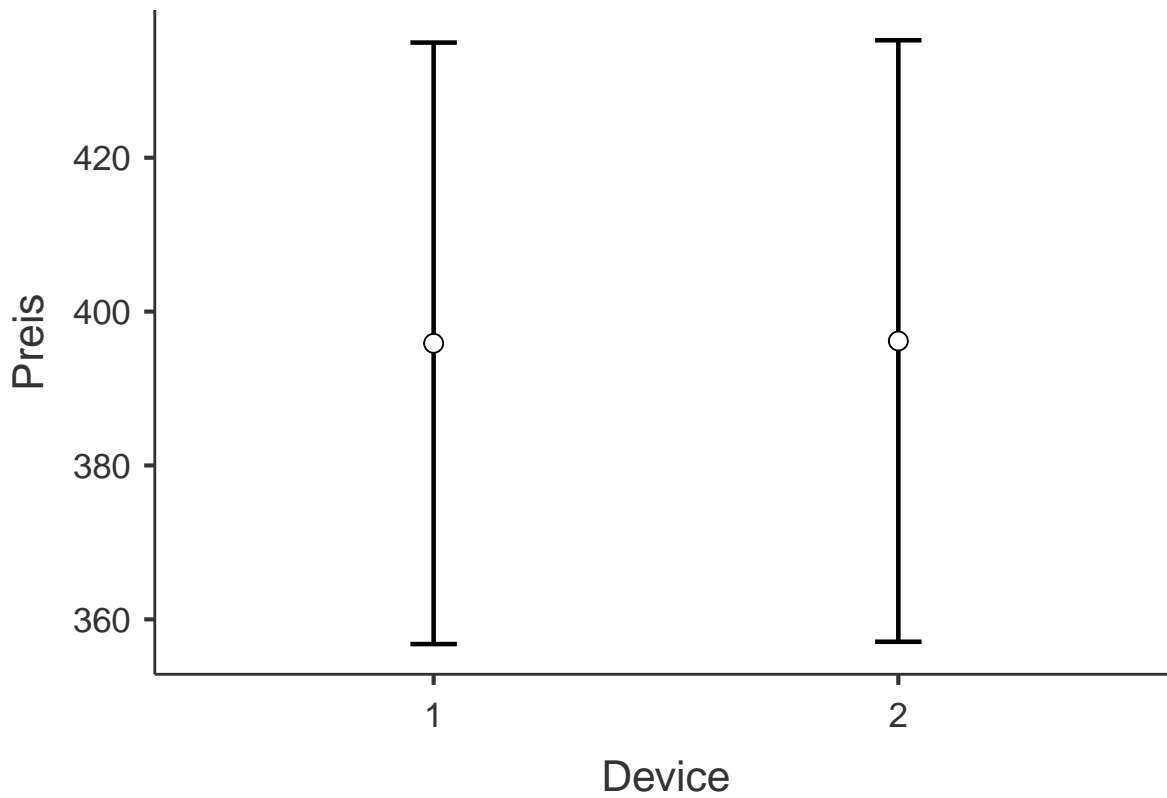
Verteilung - BoxPlot



Einfaktorielle ANOVA:

```
##
## ANOVA
##
## ANOVA
## -----
##              Sum of Squares      df      Mean Square      F      p
## -----
##      Device              12.9        1           12.9      1.16e-4    0.991
##      Residuals        6.19e+7      558        110897.6
## -----
##
##
## ESTIMATED MARGINAL MEANS
##
```

```
## DEVICE
##
## Estimated Marginal Means - Device
## -----
##      Device      Mean      SE      Lower      Upper
## -----
##      1          396      19.9       357       435
##      2          396      19.9       357       435
## -----
```



Interpretation - H3

Die ANOVA bestätigt, dass es keine signifikante Unterschiede zwischen den Gruppen gibt. Das heißt, die Hypothese 3 wird verworfen.

Hypothese 4

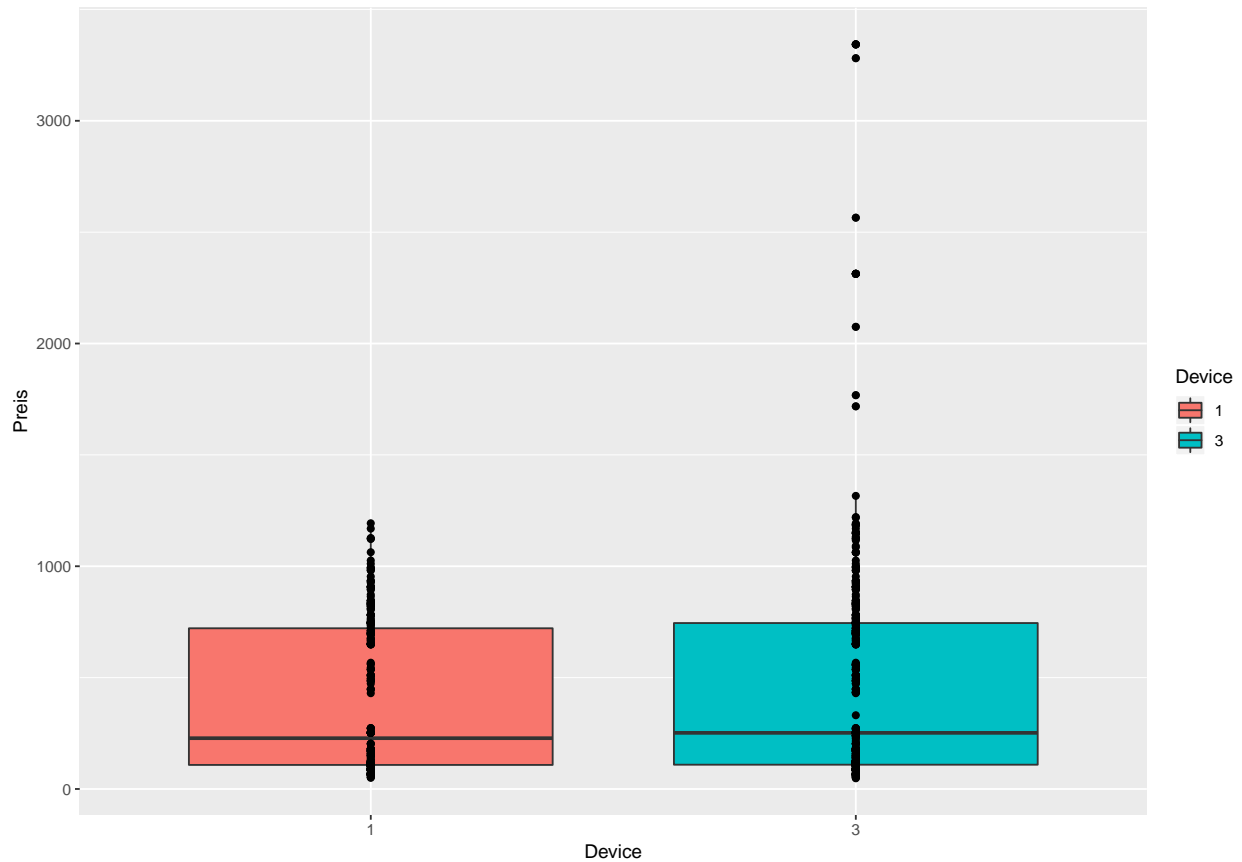
[Hypothese 4: Das Abrufen einer Reise-Website mittels Applikation und Website erwirkt einen Unterschied des offerierten Preises einer Airline.]

Deskriptive Statistik:

```
##
## DESCRIPTIVES
##
## Descriptives
## -----
```

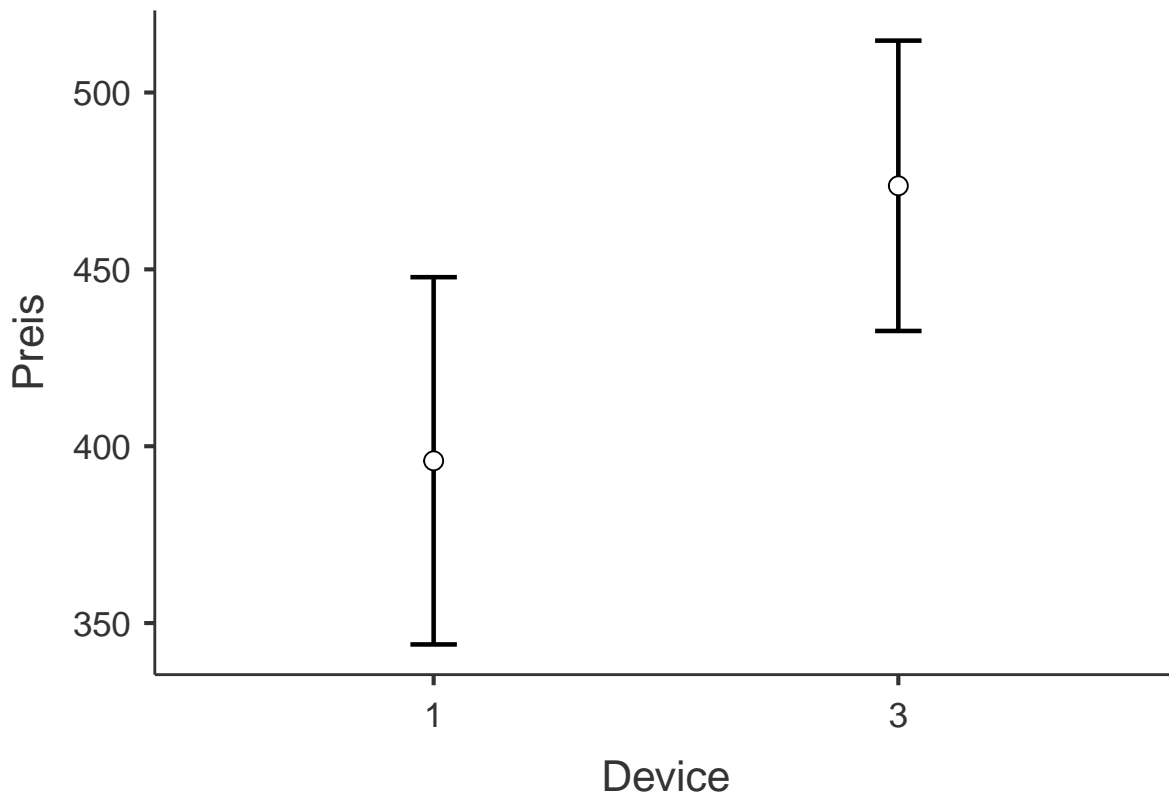
```
##                               Preis    Device
## -----
##      N                          728      728
##    Missing                       0        0
##      Mean                       444
##      Median                      252
##    Standard deviation            444
##      Minimum                     48.0
##      Maximum                     3343
## -----
##
##
##
## FREQUENCIES
##
## Frequencies of Device
## -----
##      Levels    Counts    % of Total    Cumulative %
## -----
##      1         280      38.5         38.5
##      3         448      61.5         100.0
## -----
```

Verteilung - BoxPlot



Einfaktorielle ANOVA:

```
##
## ANOVA
##
## ANOVA
## -----
##              Sum of Squares    df    Mean Square    F      p      <U+03B7>²p
## -----
## Device              1041730      1      1041730    5.32   0.021   0.007
## Residuals          1.42e+8    726      195789
## -----
##
##
## ESTIMATED MARGINAL MEANS
##
## DEVICE
##
## Estimated Marginal Means - Device
## -----
## Device    Mean    SE    Lower    Upper
## -----
## 1          396    26.4    344     448
## 3          474    20.9    433     515
## -----
```



Interpretation - H4

Es lässt sich ein signifikanter Unterschied im Preis zwischen den zwei Gruppen feststellen. Der Emm-Table zeigt, dass der höhere Preis bei Mac-Books (3) angeboten wird. Es besteht Beweis für die Hypothese 4. Diese wird daher beibehalten.

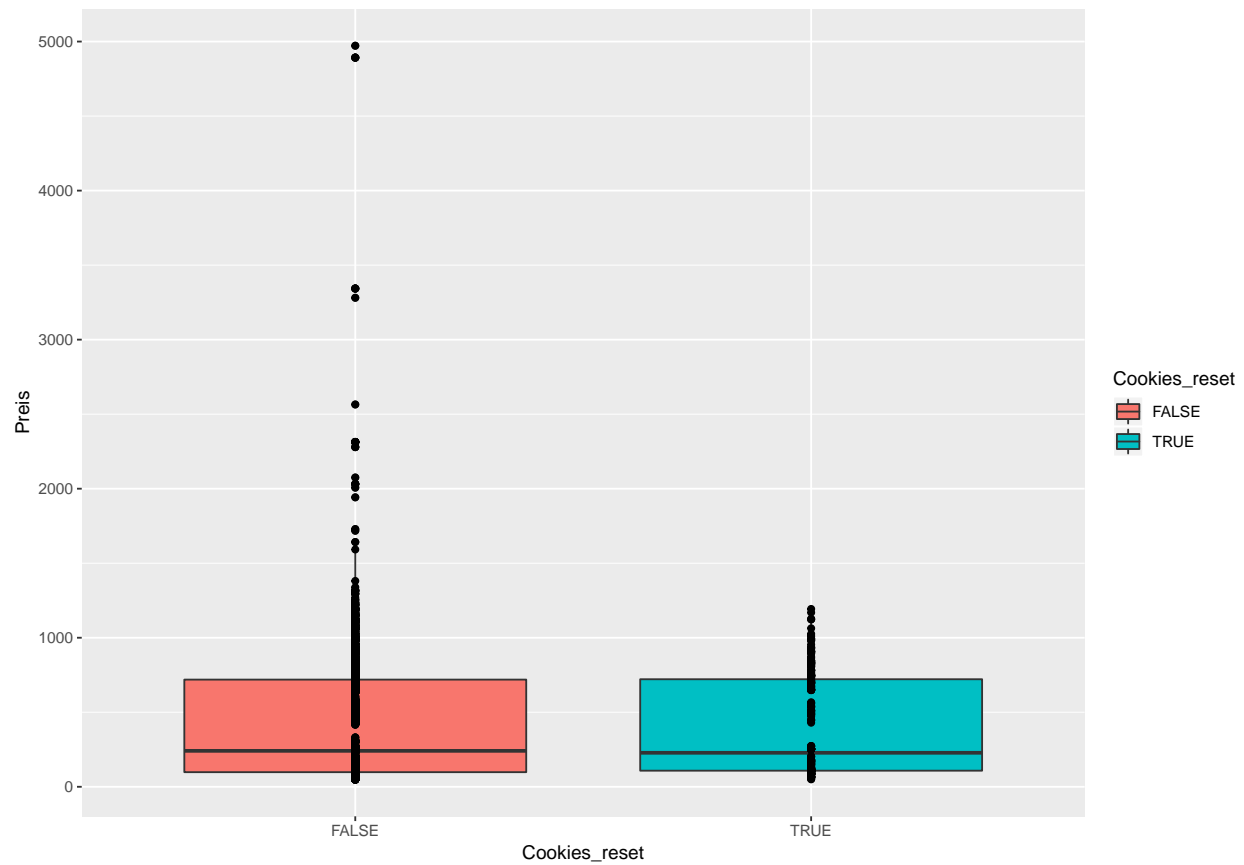
Hypothese 6

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

Deskreptive Statistik:

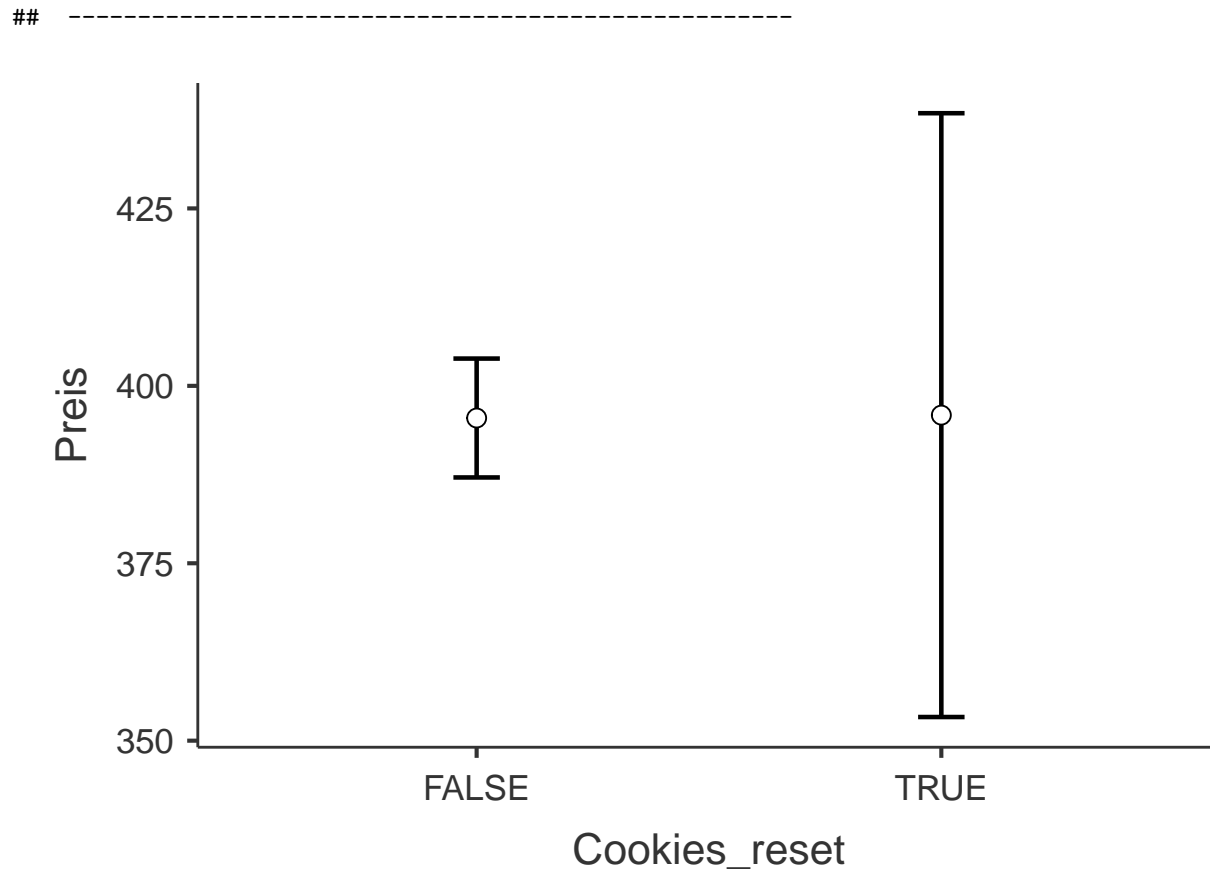
```
##
## DESCRIPTIVES
##
## Descriptives
## -----
##               Preis      Cookies_reset
## -----
##      N                7504            7504
##      Missing              0              0
##      Mean                395
##      Median              241
##      Standard deviation   363
##      Minimum             48.0
##      Maximum            4972
## -----
##
##
## FREQUENCIES
##
## Frequencies of Cookies_reset
## -----
##      Levels      Counts      % of Total      Cumulative %
## -----
##      FALSE       7224         96.3           96.3
##      TRUE         280          3.7           100.0
## -----
```

Verteilung - BoxPlot



Einfaktorielle ANOVA:

```
##
## ANOVA
##
## ANOVA
## -----
##               Sum of Squares    df    Mean Square    F        p        <U+03B7>2p
## -----
## Cookies_reset           44.9         1         44.9    3.40e-4    0.985    0.000
## Residuals          9.89e+8      7502    131872.5
## -----
##
##
## ESTIMATED MARGINAL MEANS
##
## COOKIES_RESET
##
## Estimated Marginal Means - Cookies_reset
## -----
## Cookies_reset    Mean    SE      Lower    Upper
## -----
## FALSE           395     4.27     387     404
## TRUE            396    21.70     353     438
```



Interpretation - H6

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 beizubehalten. Das heißt, es gibt einen signifikanten Unterschied zwischen den zwei Gruppen, *ceteris paribus*.

Hypothese 7

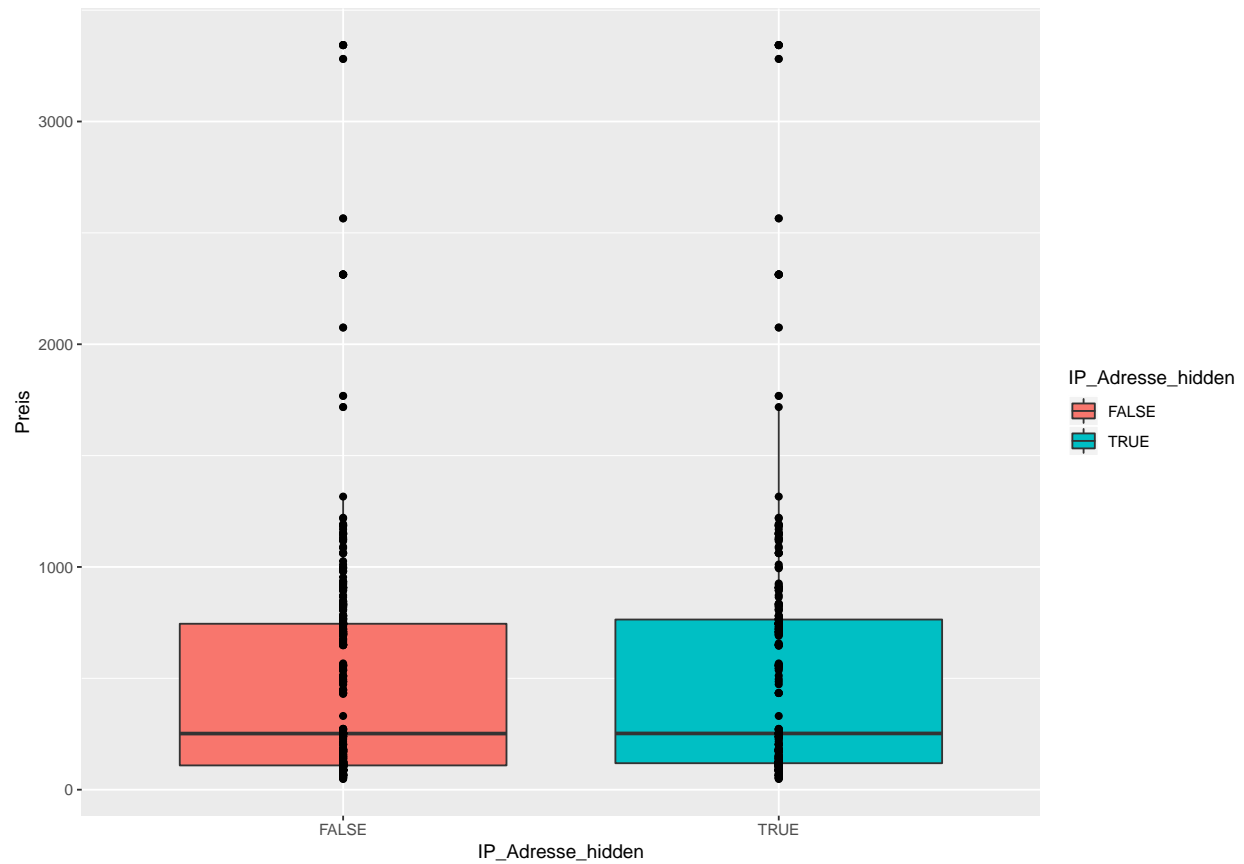
[Das Verbergen der Internetprotokoll-Adresse und folglich der ortsspezifischen Parameter mittels Virtual Private Network verursacht eine Differenz im offerierten Preis einer Airline.]

Deskriptive Statistik:

```
##
## DESCRIPTIVES
##
## Descriptives
## -----
##               Preis      IP_Adresse_hidden
## -----
##      N              728              728
##      Missing          0              0
##      Mean            496
##      Median          252
```

```
## Standard deviation      529
## Minimum                48.0
## Maximum                3343
## -----
##
##
## FREQUENCIES
##
## Frequencies of IP_Adresse_hidden
## -----
## Levels    Counts    % of Total    Cumulative %
## -----
## FALSE      448      61.5      61.5
## TRUE       280      38.5      100.0
## -----
```

Verteilung - BoxPlot



Einfaktorielle ANOVA:

```
##
## ANOVA
##
## ANOVA
## -----
## Sum of Squares    df    Mean Square    F    p    <U+03B7>2p
```

```
## -----
##   IP_Adresse_hidden      577557      1      577557      2.07      0.151      0.003
##   Residuals              2.03e+8     726      279644
## -----
##
## ESTIMATED MARGINAL MEANS
##
## IP_ADRESSE_HIDDEN
##
## Estimated Marginal Means - IP_Adresse_hidden
## -----
##   IP_Adresse_hidden      Mean      SE      Lower      Upper
## -----
##   FALSE                474      25.0      425      523
##   TRUE                  532      31.6      469      594
## -----
```



Interpretation - H7

Es gibt eine 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 beizubehalten. Das heißt, es gibt keinen signifikanten Unterschied im Mittelwert der zwei Gruppen. Bei Betrachtung des Emm ist ersichtlich, dass das **Verbergen** der Cookies mit einem vermutlich höheren mittleren Preis verbunden ist.

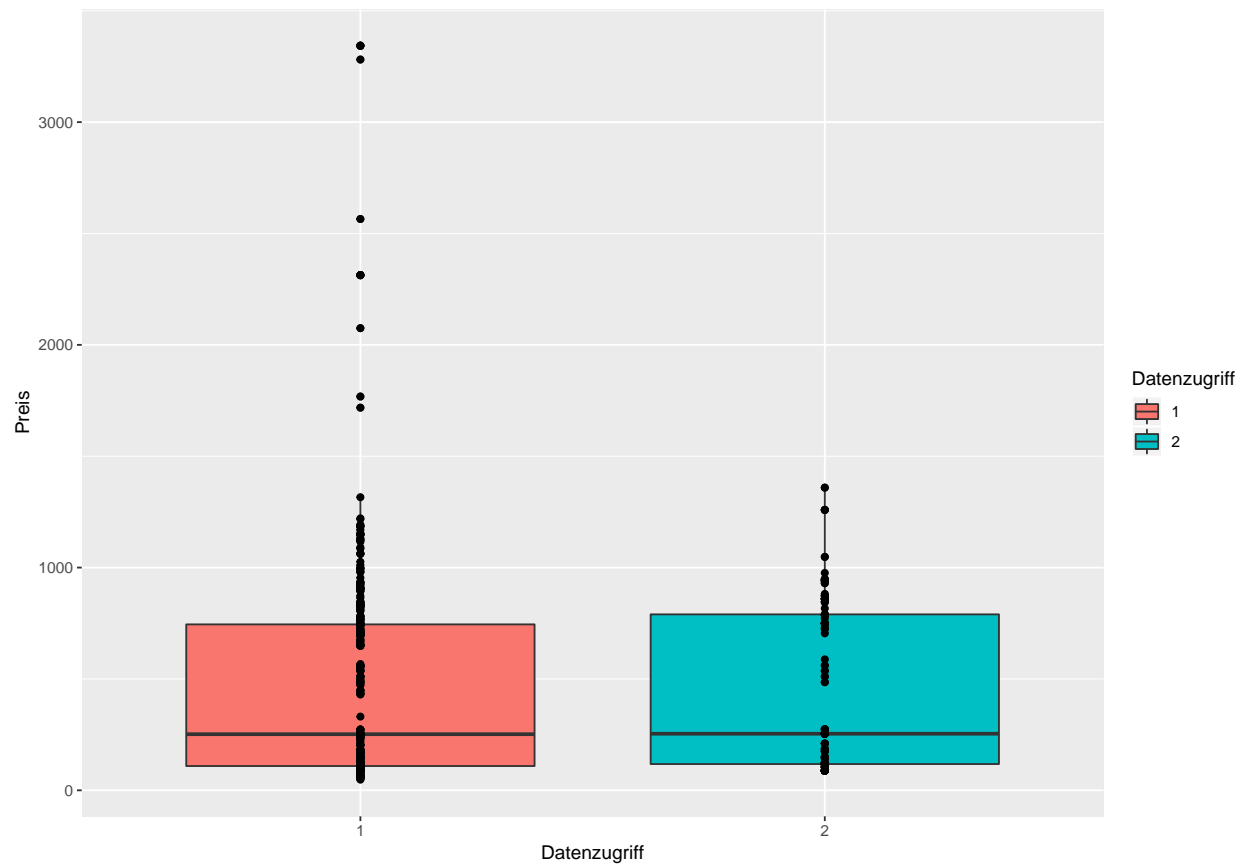
Hypothese 8

[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                588            588
##  Missing           0              0
##  Mean             468
##  Median           253
##  Standard deviation 471
##  Minimum          48.0
##  Maximum          3343
##  -----
##
##
##  FREQUENCIES
##
##  Frequencies of Datenzugriff
##  -----
##  Levels    Counts    % of Total    Cumulative %
##  -----
##  1          448       76.2         76.2
##  2          140       23.8         100.0
##  -----
```

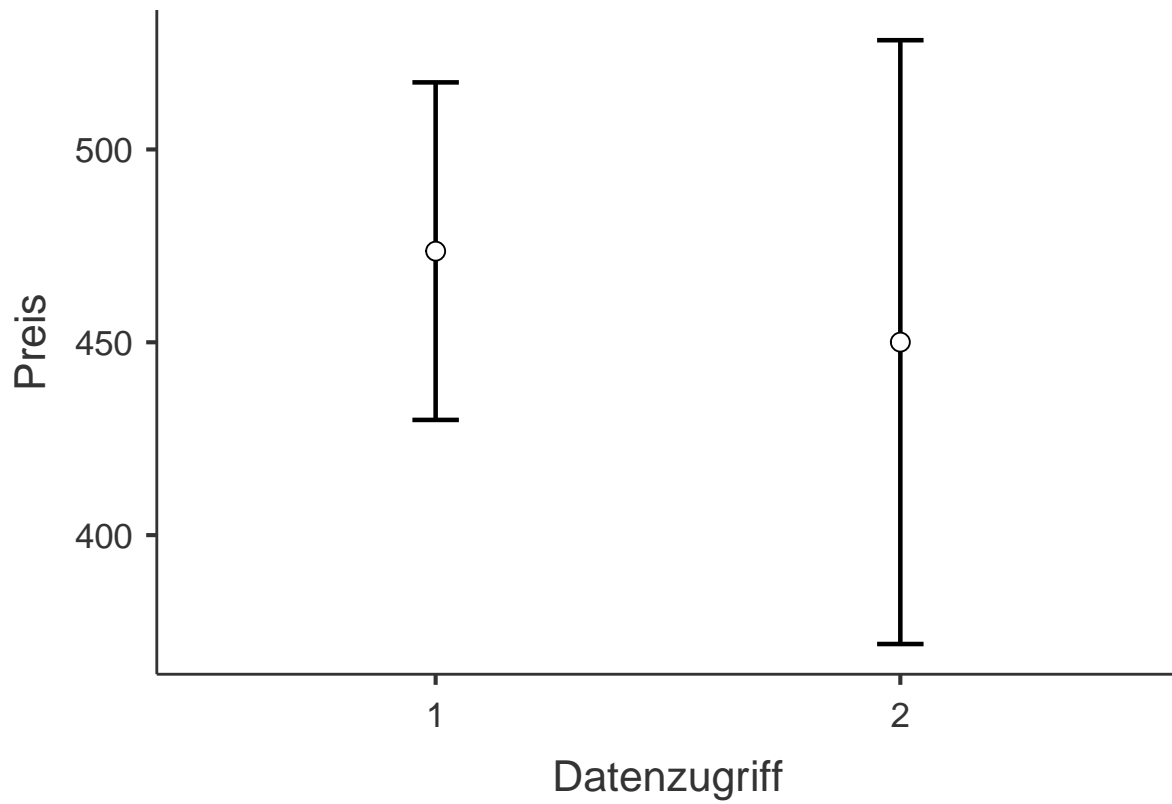
Verteilung - BoxPlot



Einfaktorielle ANOVA:

```
##
## ANOVA
##
## ANOVA
## -----
##              Sum of Squares    df    Mean Square    F      p      <U+03B7>²p
## -----
##   Datenzugriff           59342      1         59342    0.267  0.606  0.000
##   Residuals          1.30e+8    586         222431
## -----
##
##
## ESTIMATED MARGINAL MEANS
##
## DATENZUGRIFF
##
## Estimated Marginal Means - Datenzugriff
## -----
##   Datenzugriff    Mean    SE    Lower    Upper
## -----
##   1              474    22.3    430    517
##   2              450    39.9    372    528
```


##



Interpretation - H8

Die Unterschiede sind nicht signifikant. 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 Disbalancen zwischen den Sample-Größen der Merkmale.