Master

Hier wird 01\_Daten einlesen.rmd eingebunden

Daten werden im Hintergrund hier eingelesen.

#Anzeigen, welche Variablen in den Datensätzen sind (ohne weitere Information , nur Name)   
names(umsatzdaten)

## [1] "Datum" "Warengruppe" "Umsatz" "Wochentag"

names(kiwo)

## [1] "Datum" "KielerWoche"

names(wetter)

## [1] "Datum" "Bewoelkung" "Temperatur"   
## [4] "Windgeschwindigkeit" "Wettercode"

#Mehr Informationen zu den Datensätzen anzeigen  
skim(umsatzdaten)

Data summary

|  |  |
| --- | --- |
| Name | umsatzdaten |
| Number of rows | 10899 |
| Number of columns | 4 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| Date | 1 |
| factor | 1 |
| numeric | 2 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: Date**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| skim\_variable | n\_missing | complete\_rate | min | max | median | n\_unique |
| Datum | 0 | 1 | 2013-07-01 | 2019-06-06 | 2016-06-17 | 2121 |

**Variable type: factor**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| skim\_variable | n\_missing | complete\_rate | ordered | n\_unique | top\_counts |
| Wochentag | 0 | 1 | FALSE | 7 | Son: 1570, Die: 1568, Sam: 1564, Mon: 1562 |

**Variable type: numeric**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
| Warengruppe | 0 | 1 | 3.09 | 1.49 | 1.00 | 2.00 | 3.0 | 4.00 | 6.00 | ▇▃▃▃▁ |
| Umsatz | 0 | 1 | 206.66 | 142.81 | 7.05 | 97.53 | 163.3 | 280.81 | 1879.46 | ▇▁▁▁▁ |

skim(kiwo)

Data summary

|  |  |
| --- | --- |
| Name | kiwo |
| Number of rows | 72 |
| Number of columns | 2 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| Date | 1 |
| numeric | 1 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: Date**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| skim\_variable | n\_missing | complete\_rate | min | max | median | n\_unique |
| Datum | 0 | 1 | 2012-06-16 | 2019-06-30 | 2015-12-23 | 72 |

**Variable type: numeric**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
| KielerWoche | 0 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | ▁▁▇▁▁ |

skim(wetter)

Data summary

|  |  |
| --- | --- |
| Name | wetter |
| Number of rows | 2601 |
| Number of columns | 5 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| Date | 1 |
| numeric | 4 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: Date**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| skim\_variable | n\_missing | complete\_rate | min | max | median | n\_unique |
| Datum | 0 | 1 | 2012-01-01 | 2019-08-01 | 2015-12-24 | 2601 |

**Variable type: numeric**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
| Bewoelkung | 10 | 1.00 | 4.81 | 2.58 | 0.00 | 3.00 | 6 | 7.00 | 8.00 | ▃▂▂▆▇ |
| Temperatur | 0 | 1.00 | 12.10 | 7.17 | -10.25 | 6.51 | 12 | 17.82 | 32.67 | ▁▆▇▇▁ |
| Windgeschwindigkeit | 0 | 1.00 | 11.19 | 4.12 | 3.00 | 8.00 | 10 | 14.00 | 35.00 | ▆▇▂▁▁ |
| Wettercode | 669 | 0.74 | 37.48 | 27.57 | 0.00 | 10.00 | 28 | 61.00 | 95.00 | ▆▃▁▇▁ |

<<<<<<< HEAD

#führt Umsatzdaten und Kieler Woche zusammen. Wo keine Werte für kiwo, werden N/A s erzeugt  
daten\_umsatzdaten\_kiwo <- left\_join(umsatzdaten, kiwo)

## Joining, by = "Datum"

#führt Kieler Woche und Wetter zusammen  
daten\_kiwo\_wetter <- left\_join(kiwo, wetter)

## Joining, by = "Datum"

#führt Umsatzdaten. Kieler Woche und Wetter zusammen  
daten\_umsatzdaten\_kiwo\_wetter <- left\_join(daten\_umsatzdaten\_kiwo, daten\_kiwo\_wetter)

## Joining, by = c("Datum", "KielerWoche")

=======

Joinen von Daten

umsatzdaten\_kiwo<-left\_join(umsatzdaten,kiwo, by="Datum")  
names(umsatzdaten\_kiwo)

## [1] "Datum" "Warengruppe" "Umsatz" "Wochentag" "KielerWoche"

umsatzdaten\_kiwo\_wetter<-left\_join(umsatzdaten\_kiwo , wetter, by="Datum")  
names(umsatzdaten\_kiwo\_wetter)

## [1] "Datum" "Warengruppe" "Umsatz"   
## [4] "Wochentag" "KielerWoche" "Bewoelkung"   
## [7] "Temperatur" "Windgeschwindigkeit" "Wettercode"

write.csv(umsatzdaten\_kiwo\_wetter,file.path('Daten','umsatzdaten\_kiwo\_wetter.csv',fsep = .Platform$file.sep))

3ecac330827a418d93e57afece716347ebb860be

Hier wird 02\_Data\_cleaning.rmd eingebunden

Delete all rows with n/a in Wettercode

umsatzWetter <- left\_join(umsatzdaten,wetter, by="Datum")  
umsatzWetter <- umsatzWetter %>% drop\_na(Wettercode)

Hier wird 03\_Model.rmd eingebunden

#Multiple linear regression Umsatz and Temperatur, Bewölkung, Windgeschwindigkeit  
umsatz <- umsatzWetter[["Umsatz"]]  
temperatur <- umsatzWetter[["Temperatur"]]  
bewoelkung <- umsatzWetter[["Bewoelkung"]]  
windgeschwindigkeit <- umsatzWetter[["Windgeschwindigkeit"]]  
model <- lm(umsatz ~ temperatur+bewoelkung+windgeschwindigkeit)  
summary(model)

##   
## Call:  
## lm(formula = umsatz ~ temperatur + bewoelkung + windgeschwindigkeit)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -228.26 -96.40 -40.09 75.15 1709.26   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 147.6089 6.1265 24.093 <2e-16 \*\*\*  
## temperatur 4.4919 0.2314 19.413 <2e-16 \*\*\*  
## bewoelkung -0.2258 0.6347 -0.356 0.7220   
## windgeschwindigkeit 0.5670 0.3440 1.648 0.0994 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 135.3 on 8293 degrees of freedom  
## Multiple R-squared: 0.0486, Adjusted R-squared: 0.04825   
## F-statistic: 141.2 on 3 and 8293 DF, p-value: < 2.2e-16

Alternative von Anke

#data<-read\_csv(file.path('Daten','umsatzdaten\_kiwo\_wetter.csv',fsep = .Platform$file.sep))  
data<-umsatzdaten\_kiwo\_wetter  
#Multiple linear regression Umsatz vorhergesagt durch Temperatur und Bewölkung  
#definiere Model  
M1<-Umsatz ~ Temperatur+Bewoelkung  
#führe Regression durch  
model\_M1 <- lm(M1,data)  
summary(model\_M1)

##   
## Call:  
## lm(formula = M1, data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -233.80 -99.97 -40.51 76.71 1706.80   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 157.6872 4.3794 36.007 <2e-16 \*\*\*  
## Temperatur 4.3309 0.2032 21.311 <2e-16 \*\*\*  
## Bewoelkung -0.4585 0.5496 -0.834 0.404   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 139.5 on 10761 degrees of freedom  
## (135 observations deleted due to missingness)  
## Multiple R-squared: 0.04865, Adjusted R-squared: 0.04847   
## F-statistic: 275.2 on 2 and 10761 DF, p-value: < 2.2e-16

# Interpretation: Bewölkung hat keine Vorhersagewert für Umsatz über Temperatur hinaus  
  
#Aufnahme von Windgeschwindigkeit als zusätzlichen Prädiktor  
M2<-update(M1,.~.+Windgeschwindigkeit)  
model\_M2 <- lm(M2,data)  
summary(model\_M2)

##   
## Call:  
## lm(formula = M2, data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -233.66 -99.91 -40.78 76.21 1707.36   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 152.1883 5.5515 27.414 <2e-16 \*\*\*  
## Temperatur 4.3248 0.2032 21.279 <2e-16 \*\*\*  
## Bewoelkung -0.5003 0.5502 -0.909 0.363   
## Windgeschwindigkeit 0.5239 0.3251 1.612 0.107   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 139.5 on 10760 degrees of freedom  
## (135 observations deleted due to missingness)  
## Multiple R-squared: 0.04888, Adjusted R-squared: 0.04862   
## F-statistic: 184.3 on 3 and 10760 DF, p-value: < 2.2e-16

#Aufnahme von Wochentag als zusätzlichen Prädiktor  
#Dummy-Variablen, Montag ist Referenz (=0)  
M3<-update(M1,.~.+Windgeschwindigkeit+Wochentag)  
model\_M3 <- lm(M3,data)  
summary(model\_M3)

##   
## Call:  
## lm(formula = M3, data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -266.70 -95.75 -39.09 82.04 1720.58   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 131.4326 6.3979 20.543 <2e-16 \*\*\*  
## Temperatur 4.3679 0.2003 21.807 <2e-16 \*\*\*  
## Bewoelkung -0.4150 0.5426 -0.765 0.4443   
## Windgeschwindigkeit 0.5811 0.3204 1.813 0.0698 .   
## WochentagDienstag 6.2836 4.9384 1.272 0.2033   
## WochentagMittwoch 8.5981 4.9526 1.736 0.0826 .   
## WochentagDonnerstag -0.7545 4.9546 -0.152 0.8790   
## WochentagFreitag 6.5520 4.9494 1.324 0.1856   
## WochentagSamstag 54.7845 4.9458 11.077 <2e-16 \*\*\*  
## WochentagSonntag 59.0220 4.9463 11.933 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 137.4 on 10754 degrees of freedom  
## (135 observations deleted due to missingness)  
## Multiple R-squared: 0.07724, Adjusted R-squared: 0.07647   
## F-statistic: 100 on 9 and 10754 DF, p-value: < 2.2e-16

#Interpretation: samstags und sonntags wird im Vergleich zu montags sig. mehr Umsatz erzielt, anderen Tage unterscheiden sich auf 5% Niveau nicht von montags (Mittwoch etwas mehr, aber nur auf 10% sig, freitags einziger Wochentag, an dem noch weniger als montags verkauft wird)  
  
  
  
#Modellvergleich  
anova(model\_M1,model\_M2,model\_M3)

## Analysis of Variance Table  
##   
## Model 1: Umsatz ~ Temperatur + Bewoelkung  
## Model 2: Umsatz ~ Temperatur + Bewoelkung + Windgeschwindigkeit  
## Model 3: Umsatz ~ Temperatur + Bewoelkung + Windgeschwindigkeit + Wochentag  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 10761 209345155   
## 2 10760 209294638 1 50517 2.6754 0.1019   
## 3 10754 203054707 6 6239931 55.0789 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

#Interpretation: Aufnahme von Windgeschwindigkeit erhöht nicht die Modellpassung, Aufnahme von Wochentags-Dummies wohl