HW3

INTRODUCTION:

In this homework my aim is to build models to forecast daily production the same as project and to compare them. I built 3 models which are ARIMA, SARIMA and SARIMA model with DSWRF values. I used auto.arima functions to find the parameters of the models.

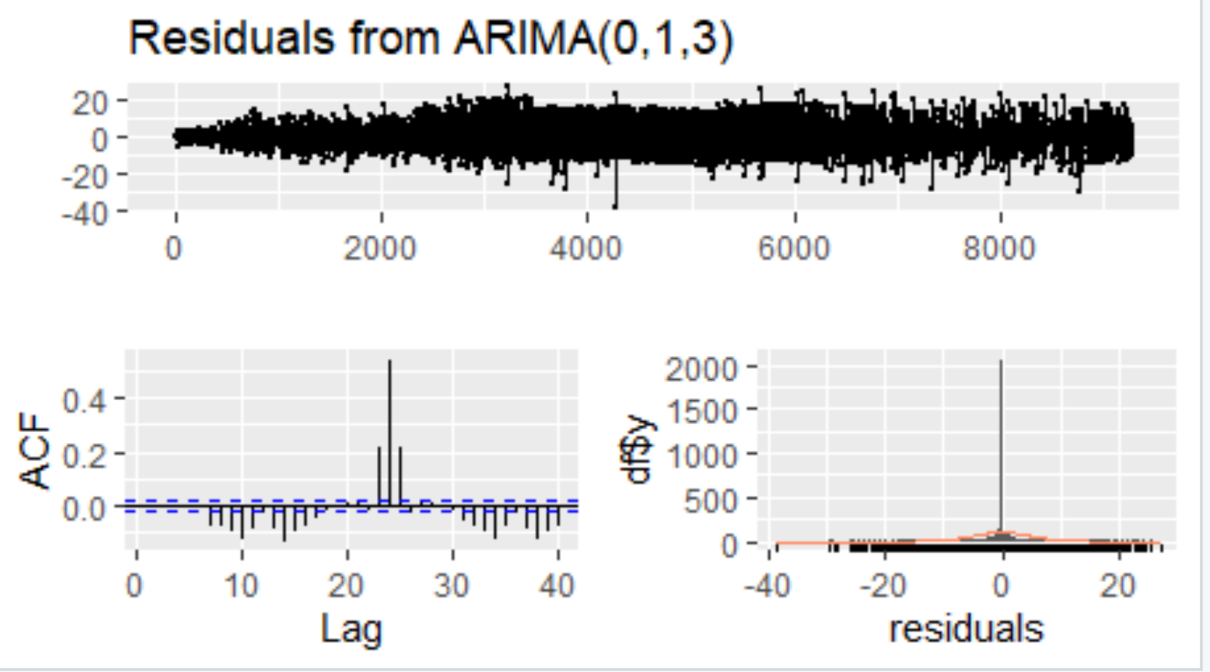
Preparing the data:

```
library(RcppRoll)
library(tidyr)
library(readxl)
library(lubridate)
library(zoo)
library(ggplot2)
library(scales)
library(data.table)
library(corrplot)
library(ggcorrplot)
library(GGally)
library(forecast)
library(dplyr)

production=fread("2022-06-13_production.csv")
weather=fread("2022-06-13_weather.csv")


production = production[order(date,hour)]
production = production[,month:= as.factor(month(date))]
production = production[,quart:= as.factor(quarter(date))]
head(production,3)

wide_weather=dcast(weather,date+hour~variable+lat+lon,value.var='value')

production_with_weather=merge(production,wide_weather,by=c('date','hour'))
train_data=production_with_weather[date<'2022-03-01']
test_data=production_with_weather[date>='2022-03-01' & date<='2022-05-24']
tail(train_data[,1:6],3)
head(test_data[,1:6],3)
tail(test_data[,1:6],2)
nrow(train_data)
nrow(test_data)
```

ARIMA model:

```
arima_model =
auto.arima(train_data$production,seasonal=F,trace=T,stepwise=T,approximation=T)
arima_model
```

```
Fitting models using approximations to speed things up...

 ARIMA(2,1,2) with drift         : 55557.13
 ARIMA(0,1,0) with drift         : 58555.86
 ARIMA(1,1,0) with drift         : 55863.46
 ARIMA(0,1,1) with drift         : 55753.76
 ARIMA(0,1,0)                    : 58553.85
 ARIMA(1,1,2) with drift         : 55554.12
 ARIMA(0,1,2) with drift         : 55553.35
 ARIMA(0,1,3) with drift         : 55553.17
 ARIMA(1,1,3) with drift         : 55556.25
 ARIMA(0,1,4) with drift         : 55555.05
 ARIMA(1,1,4) with drift         : Inf
 ARIMA(0,1,3)                    : 55551.16
 ARIMA(0,1,2)                    : 55551.34
 ARIMA(1,1,3)                    : 55554.25
 ARIMA(0,1,4)                    : 55553.04
 ARIMA(1,1,2)                    : 55552.12
 ARIMA(1,1,4)                    : Inf

 Now re-fitting the best model(s) without approximations...

 ARIMA(0,1,3)                    : 55554.99

 Best model: ARIMA(0,1,3)

> arima_model
Series: train_data$production
ARIMA(0,1,3)

Coefficients:
         ma1     ma2     ma3
      0.5984  0.1552  0.0152
s.e.  0.0104  0.0120  0.0103

sigma^2 = 23.19:  log likelihood = -27773.49
AIC=55554.99   AICc=55554.99   BIC=55583.53



checkresiduals(arima_model)
```

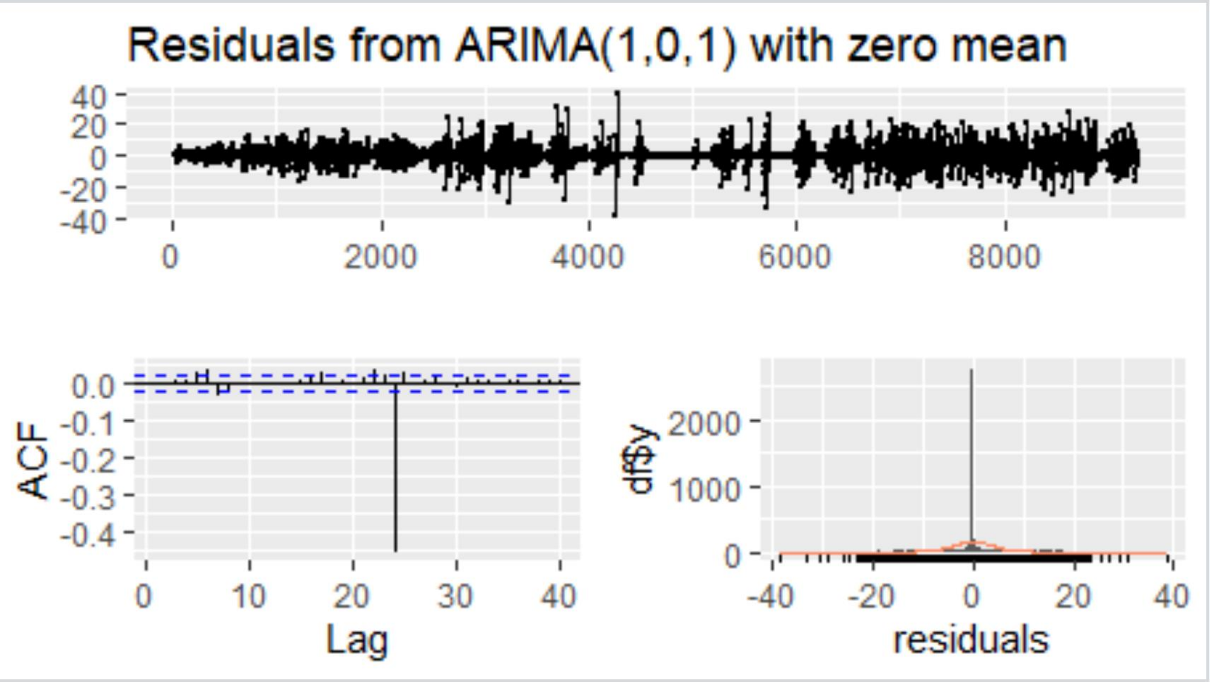


```
Ljung-Box test

data:  Residuals from ARIMA(0,1,3)
Q* = 365.8, df = 7, p-value < 2.2e-16

Model df: 3.   Total lags used: 10
```

SARIMA model:

```
sarima_model=auto.arima(train_data$diff_series,seasonal=T,trace=T,stepwise=T,approximation=T)
sarima_model
```

```
Fitting models using approximations to speed things up...

 ARIMA(2,0,2) with non-zero mean : 50518.63
 ARIMA(0,0,0) with non-zero mean : 58604.59
 ARIMA(1,0,0) with non-zero mean : 50674.02
 ARIMA(0,0,1) with non-zero mean : 53065.32
 ARIMA(0,0,0) with zero mean     : 58602.64
 ARIMA(1,0,2) with non-zero mean : 50514.93
 ARIMA(0,0,2) with non-zero mean : 51459.56
 ARIMA(1,0,1) with non-zero mean : 50514.03
 ARIMA(2,0,1) with non-zero mean : 50516.05
 ARIMA(2,0,0) with non-zero mean : 50524.87
 ARIMA(1,0,1) with zero mean     : 50512.03
 ARIMA(0,0,1) with zero mean     : 53063.35
 ARIMA(1,0,0) with zero mean     : 50672.03
 ARIMA(2,0,1) with zero mean     : 50514.06
 ARIMA(1,0,2) with zero mean     : 50512.94
 ARIMA(0,0,2) with zero mean     : 51457.58
 ARIMA(2,0,0) with zero mean     : 50522.88
 ARIMA(2,0,2) with zero mean     : 50516.63

 Now re-fitting the best model(s) without approximations...

 ARIMA(1,0,1) with zero mean     : 50511.93

 Best model: ARIMA(1,0,1) with zero mean

> sarima_model
Series: train_data$diff_series
ARIMA(1,0,1) with zero mean

Coefficients:
         ar1     ma1
      0.6847  0.1764
s.e.  0.0099  0.0135

sigma^2 = 13.65:  log likelihood = -25252.96
AIC=50511.92   AICc=50511.93   BIC=50533.33
checkresiduals(sarima_model)
```



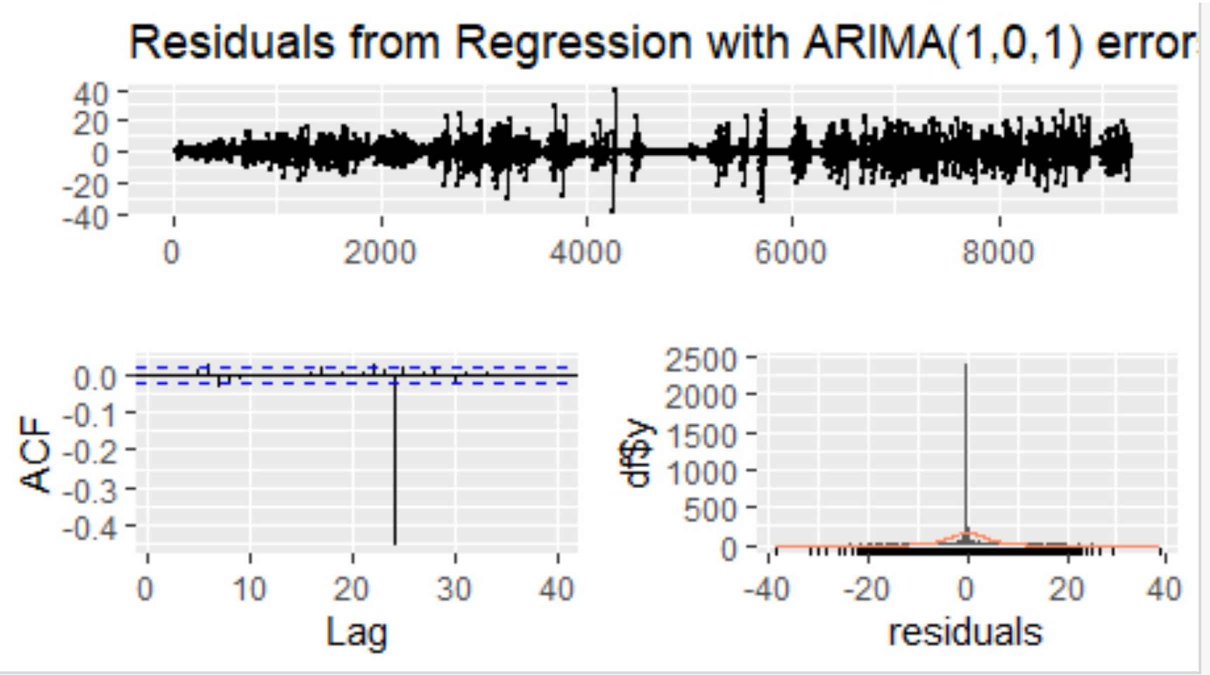
```
Ljung-Box test

data:  Residuals from ARIMA(1,0,1) with zero mean
Q* = 34.842, df = 8, p-value = 2.857e-05

Model df: 2.   Total lags used: 10
```

SARIMA model with DSWRF values:

```
sarimax_model=
auto.arima(train_data$diff_series,xreg=reg_matrix,seasonal=T,trace=T,stepwise=T,approximation=T)
sarimax_model
```

```
Fitting models using approximations to speed things up...

 ARIMA(2,0,2) with non-zero mean : 50370.4
 ARIMA(0,0,0) with non-zero mean : 57141.94
 ARIMA(1,0,0) with non-zero mean : 50534.86
 ARIMA(0,0,1) with non-zero mean : 52257.75
 ARIMA(0,0,0) with zero mean     : 57140.03
 ARIMA(1,0,2) with non-zero mean : 50367.39
 ARIMA(0,0,2) with non-zero mean : 51011.01
 ARIMA(1,0,1) with non-zero mean : 50365.59
 ARIMA(2,0,1) with non-zero mean : 50368.39
 ARIMA(2,0,0) with non-zero mean : 50374.62
 ARIMA(1,0,1) with zero mean     : 50363.6
 ARIMA(0,0,1) with zero mean     : 52255.79
 ARIMA(1,0,0) with zero mean     : 50532.87
 ARIMA(2,0,1) with zero mean     : 50366.4
 ARIMA(1,0,2) with zero mean     : 50365.4
 ARIMA(0,0,2) with zero mean     : 51009.03
 ARIMA(2,0,0) with zero mean     : 50372.63
 ARIMA(2,0,2) with zero mean     : 50368.42

 Now re-fitting the best model(s) without approximations...

 ARIMA(1,0,1) with zero mean     : 50363.4

 Best model: Regression with ARIMA(1,0,1) errors

> sarimax_model
Series: train_data$diff_series 
Regression with ARIMA(1,0,1) errors 

Coefficients:
         ar1     ma1    xreg
      0.6431  0.1871  0.0128
s.e.  0.0112  0.0139  0.0010

sigma^2 = 13.44:  log likelihood = -25177.7
AIC=50363.4   AICc=50363.4   BIC=50391.93
```

```
checkresiduals(sarimax_model)
```



```
Ljung-Box test

data:  Residuals from Regression with ARIMA(1,0,1) errors
Q* = 25.186, df = 7, p-value = 0.0007033

Model df: 3.   Total lags used: 10
```

CONCLUSION:

First, I prepare data in order to make data useful for constructing model. I do Ljung box test in order to decide data is correlated or not. Then I checked all models' residuals and plot them. After building the methods and calculating the performance measures, the best model is SARIMA model with using DSWRF values. I choose this model based on AIC, AICc, BIC and p values. Because SARIMA model with using DSWRF values has lowest AIC, AICc, BIC values I chose it.

In summary I can conclude that adding DSWRF values as a regressor improved my model because production and DSWRF values are highly correlated.