Rikards Larsson (Rilr20) Matmod Project

Uppgift 1:Beskriv data

Datan kommer ifrån SMHIs väderstationer. Väderstationerna som jag har valt är aktiva väderstationer. De stationerna som jag valde heter Malmö A, Lund, Skillinge A. Simrishamns väderstation var inte aktiv så då valdes Skillinge A istället för den var närmst och aktiv.

Väderstationen Malmö A stationen ligger inom storstadsregionen och Lunds mätstation ligger inom deras stadsgränser. Skillinge A är en mätstation som ligger vid kusten. Dessa tre stationer valdes för de är inom samma region och för att se om det finns någon differens på mätdatan.

De tre väderstationerna samlar datan olika ofta för att det ska bli en rättvis jämförelse så kommer min temperaturavläsning att ske vid 6:00 och 18:00. De resterande klockslagen ignoreras ifall det förekommer mättemperaturer.

Variablerna som används är lufttemperatur, datum, och tid. Temperaturen som mäts är i celcius. Mättillfällena är från 2 augusti till 10 december 2021. Då får man temperaturen för höst- och vinterväder i 130 dagar.

Vid vissa tidpunkter så finns det felkoder där temperaturerna avläses men jag vet inte om avläsningen har skett korrekt eftersom på en avläsning står det "stationen eller givaren har varit ur funktion.". Men ett värde har ändå skrivits in, av datan som används så har det bara hänt i Lund.

Exempel på datan som finns och används i programmet.

Mätstation	Temperatur (Celcius)	Tid	Datum
Malmö	11.6	6:00	2021-08-24
Malmö	16.1	18:00	2021-08-24
Malmö	15.0	6:00	2021-08-25
Malmö	17.0	18:00	2021-08-25
Malmö	13.7	6:00	2021-08-26
Malmö	14.0	18:00	2021-08-26

I figur 1 så ser man temperaturskillnaden i Simrishamn är markant mindre jämfört med Malmö och Lunds. Simrishamn har även tre outliers och det är de tre dagarna som är minusgrader. -2.8 klockan på morgonen och -1.9, -2.0 på kvällen. De lägsta och högsta temperaturvärderna är under 18:00 klockslaget för Lund och Malmö. Datan för Simrishamn så är den lägsta temperaturen på morgon och den högsta på kvällen. De högsta och lägsta temperaturerna förekommer i Lund.

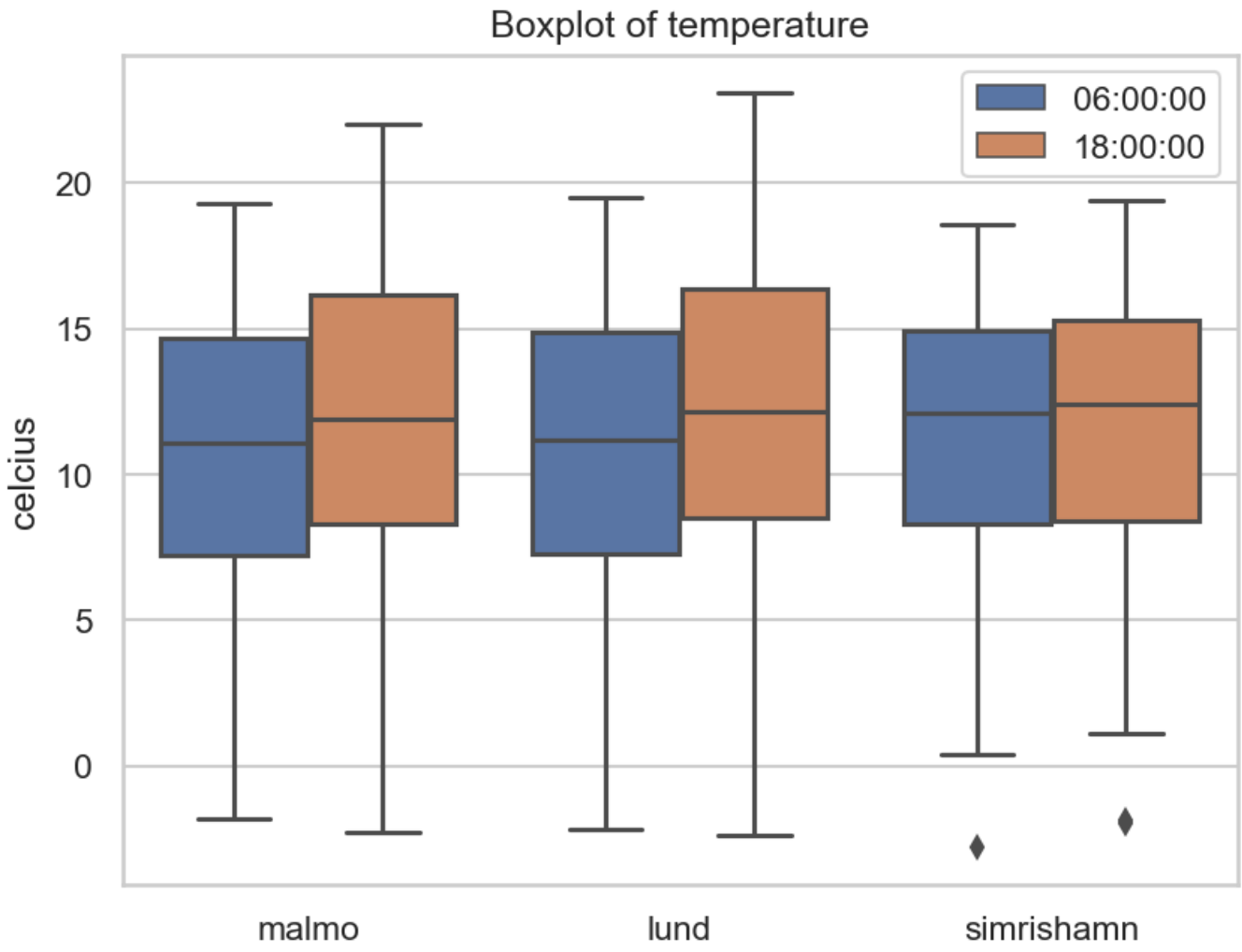

Fig 1: Lådagram över temperaturen under en 6 månaders period uppdelad i morgon- och kvällstemperatur

Uppgift 2: Beskrivande statistik

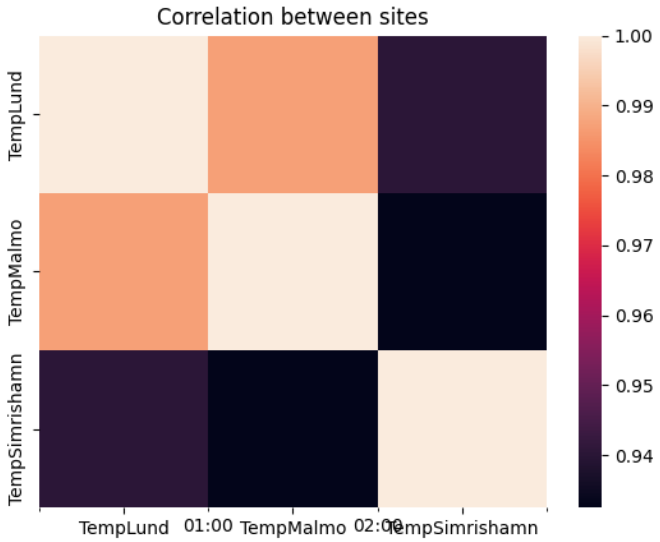

Fig 3: Korrelation mellan de tre väderstationerna

Stad	Medelvärde	Standardavvikelse	Max-värde	Min-värde
Malmö	10.970881	5.290972	22.0C	-2.3C
Simrishamn	11.483525	4.67948	19.4C	-2.8C
Lund	11.306897	5.24404	23.1C	-2.4C

Korrelationen mellan de tre väderstationerna är hög vilket kan bero på att de ligger nära till varandra. Malmö och Lunds är lika medans Simrishamn sticker ut men inte med mycket.

Simrishamn har högst medelvärde fast de har lägsta max- och min-värde. Min-värdet där var dock en outlier så resterande data kan ligga närmre varandra. Värdet på medeltemperaturen skiljer inte så mycket mellan varandra.

Uppgift 3: Beskrivande plottar

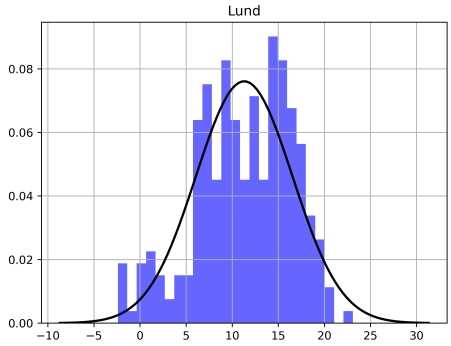

Fig 4: Histogram och normalfördelning av temperaturerna för mätstationen Lund

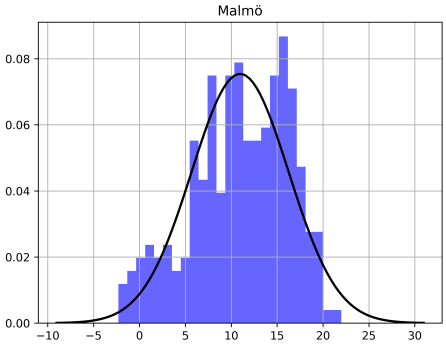

Fig 5: Histogram och normalfördelning av temperaturerna för mätstationen Malmö

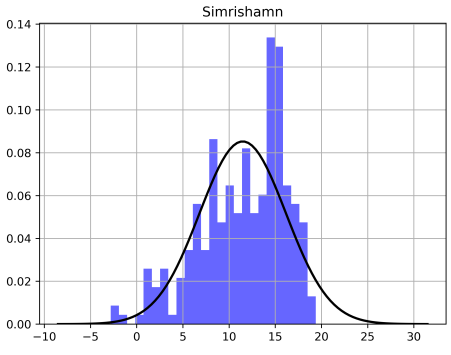

Fig 6: Histogram och normalfördelning av temperaturerna för mätstationen Simrishamn

Histogrammen ser ut att följa normalfördelningen men det skulle behövas mer data för att säkert säga att det följer. Alla tre histogrammen så är det glest med temperaturavläsningar i den övre kvantilerna. Detta kan beror på att det är senhöst väder som sedan blir vintervärde.

Simrishamn har väldigt många mätpunkter på 14 och 15 grader så att det inte förhåller sig med normalfördelningen. Detta kan vara varför medelvärdet är högre än de andra två mätstationerna.

Uppgift 4: Linjär regression

Variablerna i den linjära regressionen så är riktningskoefficienten b lika med -0.05908414 och konstanten a är lika med 18.71231311. För få ut det tvåsidiga 95 procentiga konfidensintervallet så använder man värdet 0.534335742123695. Den linjära regressionen har en negativ lutning. Denna negativa lutningen passar då det blir kallare när det går från höst till vinterväder.

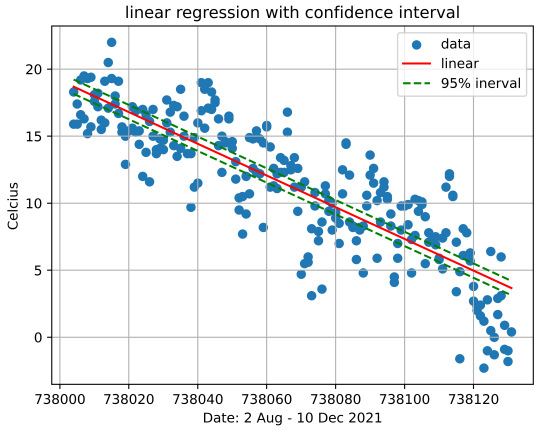

Fig 7: Linjär regression av alla tre väderstationerna

Uppgift 5: Transformerad data

Väderdatan transformerades till en logaritmisk funktion. Den logaritmiska linjen startar under de flesta mätpunkterna för temperaturen och slutar över de flesta datapunkterna i slutet av datan. Tillskillnad från den linjära regressionen som försöker vara i mitten av alla datapunkterna.

Den linjära regressionen ser ut att följa datan bättre än vad den logaritmiska funktionen gör, då datapunkterna är närmre den regressionen. Den linjära regressionen är lättare att avläsa hur temperaturen ändras sig under perioden.

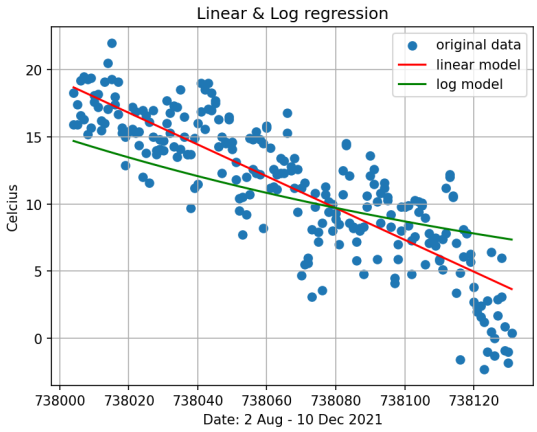

Fig 8: Linjär regression och logaritmisk regression

Uppgift 6: Residualanalys

I figur 10 ser man att residualerna för den linjära regressionen ligger relativt nära regressionen i början av datan men i slutet så är residualerna väldigt höga eller väldigt låga. När residualerna är i en normalfördelningen (fig. 9) så kan man se att standardavvikelsen inte är så hög då normalfördelningen är lång och smal. Vilket kan tyda på att den linjära regressionen passar bra då de flesta residualerna är runt 0 vilket är vad medelvärdet ligger nära.

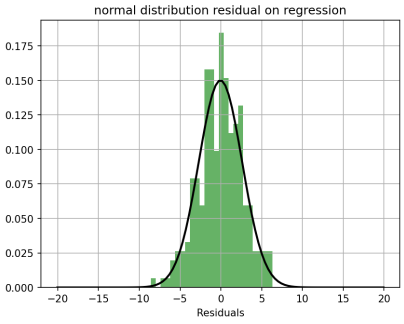

Fig 9: Linjära regressionens residualer i ett histogram och en normalfördelning med medelvärde på -0.05334

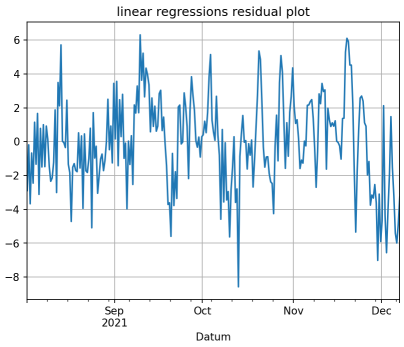

Fig 10: plot över residualerna för linjär regression

Den logaritmiska regressionens residualer (fig. 12) ser ut att följa samma mönster som med den linjära (fig. 10). Det passar bättre i början men sedan i slutet av datan så passar det inte lika bra. I normalfördelningen så är standardavvikelsen större än i figur 9. Medelvärdet är större än i den linjära regressionens normalfördelning. Detta visar att den logaritmiska regressionen passar inte lika bra som med den linjära.

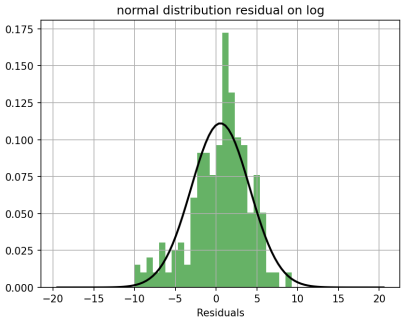

Fig 11: Logaritmisk regressionens residualer i ett histogram och en normalfördelning med medelvärde på 0.53280

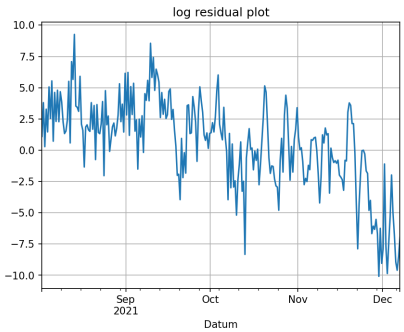

Fig 12: plot över residualerna för logaritmisk regression

Det låga temperaturerna är vad som har påverkat residualerna i slutet av figur 10 och figur 12.

Uppgift 7: Sammanfattning

Sammanfattningsvis så är den linjära regressionen den bättre regressionen för denna typ av data. Då många av residualerna ligger runt noll då medelvärdet ligger på -0.05. Datan som används är från höst till vinter påverkan detta har på datan är att regressionerna har en negativ lutning. På grund av att vädret på datan går från sensommar/höst till vinter så får man fler kallare värden än om man hade gjort denna analysen tidigare. Värderna påverkar hur histogrammen ser ut hade mätperioden varit över ett helt år så hade värderna troligtvis följt normalfördelningen bättre. Histogrammets jämförelse mot normalfördelning så ser man fler kallare värden som på den nedre kvantilerna. Under denna mätperioden har det har då blivit en överreprensentation av kallare värden.

Stationerna ligger i olika typer av miljöer detta kan påverka datan då man inte vet om stadsmiljöerna påverkar temperaturen. Det finns ingen stor differens på datan då korrelationen mellan de tre mätstationerna är relativt hög. Om mätstationerna var mer utspridda så hade korrelationen sett annorlunda ut.