Hjemmeeksamen / arbeidskrav

Eirik Heen

18-10-2022

1. Bakgrunn

I denne oppgaven skal du undersøke salg av drikkevarer ved Buktafestivalen. Bruk av datasettet i undervisning har blitt godkjent av styremedlemmer av Buktafestivalen med noen endringer (se under). Dere skal bruke datasettet til å beskrive flere sammenhenger i forbindelse med Buktafestivalen og sammenlikne år og dager. Datasettet inneholder alle salg for 4 år (2016-2019) med konserter, over de 3 dagene som festivalen varer (torsdag, fredag og lørdag).

For å kunne bruke datasettet til undervisning har all data blitt aggregert fra sekund-nivå til time- og kvarters-nivå. I tillegg har noen produkter blitt slått sammen. Dette betyr at det er mange færre produkter i datasettet enn faktisk selges på Buktafestivalen. Mye av transaksjonsdata er også fjernet. Last ned datasettet nå datasett.

Datasettet har 14 variabler og 111 114 rader. Hver rad er et salg eller deler av et salg. Hvis en kunde kjøpte 2 øl blir dette en rad hvor kvantum er 2, men hvis en kunde kjøpte 1 øl og 1 vin blir dette 2 rader hvor kvantumet er 1 i hver rad.

Variabler:

Dato - dato og tidspunkt for salg.

Antall - antall enheter solgt av **én** produkttype.

Pris - Prisen på en vare solgt.

Nedbor - Nedbør målet i mm regn per time.

Luft_temperatur - Temperatur for hver time.

Solskin - Minutter med solskinn per time.

Vind - Vindstyrke målt i meter per sekund.

Gjester - Antall gjester på området per time (ikke målt i 2017).

Produkt - Produktet som ble solgt. Nivå:

- Bukta Beer: Spesialøllet Bukta brygger til hver festival, begrenset opplag
- Pilsner: Vanlig Mack pilsner

- Other beer: Alt annet øl som selges
- Cider: Alle typer sider
- Wine: Alle typer vin
- Non alcoholic: Alle alkoholfrie drikkevarer

Ar - År salget ble registrert. Nivå: 2016, 2017, 2018, 2019.

Time - Klokkeslett for salget, kun på hele times-nivå.

Dag - Hvilken dag for salget. Nivå: Torsdag, Fredag og Lørdag

Per15min - teller antall timer fra start av konsert per 15 min. Nivå: 0, 0.25, 0.5, 0.75, osv.

2. Instruksjoner og oppgaver

Denne oppgaven er både et arbeidskrav og en gruppeeksamen. Arbeidskravsdelen av oppgaven består av muntlig presentasjon av resultatene hvor dere også kan få tilbakemelding på oppgaven. I tillegg som del av arbeidskravet skal gruppen undertegne en samarbeidsavtale.

Gruppeeksamen består av å levere inn arbeidet i ettertid i WiseFlow. Dere skal levere enn oppgave sammen og dette blir 50% av karakteren din i mappeevalueringen.

Hver gruppe kan ha fra 2 til 5 studenter og må levere inn samarbeidsavtale for gruppearbeid under den muntlige presentasjon av prosjektet i uke 44 - uke 45. På presantasjonen får dere opp til 20 min å presantere oppgaven deres, jeg blir å stille spørsmål og dere kan komme med spørsmål. For å få arbeidskravet godkjent må dere gi en presantasjon av oppgaven og lever inn samarbeidsavalen til meg. Studenter som ikke har sine navn på en samarbeidsavtale få ikke arbeidskravet godkjent. Studenter som ønsker å levere inn alene må søke administrasjonen om tillatelse for å kunne levere alene (de trenger ikke levere en samarbeidsavtale).

Dere skal levere en pdf fil som inneholder tekst, figurer og koden dere har brukt. Bruk gjerne Quarto eller rMarkdown i RStudio som plattform. I pdf filen skal tittelen være “Sok-2009, høst 2022, Gruppeeksamen”, dere skal under tittelen skrive inn kandidatnummeret til alle forfatterne i gruppen.

Dokumentet skal besvare spørsmålene nedenfor. Vi forutsetter et signifikansnivå på 5% ($\alpha = 0.05$) i alle oppgavene. Besvar oppgavene fullstendig som om du skulle gi en rapport til noen som jobbet for Buktafestivalen. Dere kan anta at de har kunnskap i statistikk, men dere må forklare resultatene grundig.

Oppgave 1. Deskriptiv statistikk, inntekter og deltakelse

Før man begynner med en statistisk analyse av et datasett er det vanlig å presentere deskriptiv statistikk. Dette gjøres for å få oversikt over de viktigste målene i et datasett, slik at du og leseren lettere kan forstå hva som kommer senere i analysen.

Lag en tabell over total inntekt per år, deltakere per år og inntekt per deltaker. Ha år på radene og nevnte variabler på kolonene. Siden vi ikke vet hvor mange deltakere som har gått inn og ut av festivalområdet skal vi for enkelhets skyld antar at totalt antall deltakere per festival er maksimalt antall registrerte deltakere på en dag. Deltakere er ikke registrert for 2017, men inkluder dette årstallet i tabellen for sammenlikning av inntekt.

Lag en tilsvarende tabell, men nå bryt samme data opp i år og dager. På denne måten kan vi sammenlikne om det er forskjell mellom år eller dagene.

Lag grafer av disse tabellene (stolpediagram), hvor du viser inntekt per år og inntekt per år og dag. Velg selv hvordan dere ønsker å presentere dataene for å få fram de forskjellene dere mener er viktig.

Fra disse to tabellene, ser det ut til å være noe forskjeller mellom årene og/eller dagene for hvert år? Vi kan ennå ikke si noe om det er statistisk signifikant, men ser du noen trender i dataene?

Oppgave 2.1 Deskriptiv statistikk, produkt type

Drikkevarene er delt inn i 6 produkttyper. Lag en tabell over total inntekt for hver produkt, og plott også resultatet i et stolpediagram. Hvilket produkt virker til å bringe inn mest inntekter og hvilket bringer inn minst?

Lag en ny tabell over total inntekt for hvert produkt *per år*. Igjen plott resultatet i et stolpediagram. Ser det ut til å være noe forskjell i distribusjonen av kjøp av drikke varer mellom forskjellige år. Altså er et produkt mer eller mindre populært noen år?

Oppgave 2.2. Statistisk testing for produkttype

I denne oppgaven antar vi at salg av produkter per 1 time er uavhengige salg, dette gjør det mulig å sammenlikne inntekt mellom år og dag. Finn total inntekt for alle produktene per time. Gjennomfør en parvis t-test mellom produktene og inntekt, korriger p-verdiene med metoden *Holm*. Hvilke produkter tjener Bukta festivalen mer eller mindre på, og hvilke er relativt lik?

Oppgave 3.1 Deskriptiv statistikk, inntekt per 15 min

Lag 3 grafer som viser sammenhengen mellom total inntekt og hver 15 min av festivalen. Lag en graf for torsdag men en linje for hvert år, en graf for fredag med en linje for hvert år og en graf for lørdag med en line for hvert år.

Hvilke trender ser vi i de forskjellige grafene og er det noe forskjell mellom år eller dager?

Oppgave 3.2. Lineær regresjon av inntekt

Vi skal sammenlikne inntekter mellom årene og dagene. For å gjøre dette skal vi gjennomføre en lineær regresjon. Aggregér datasettet ned til total inntekt per 15min, for hver dag og år (dette datasettet skal ha 240 rader). Modellen du skal kjøre står nedenfor.

$$Totalinntekt = \beta_0 + \beta_1 \text{År} + \beta_2 Dag + \beta_3 Per15min$$

Kommentér resultatene til modellen. Slik denne regresjonen er satt opp går det ikke an å si noe om forskjellig inntekt mellom fredag og lørdag, det er kun mulig å si om inntekten er forskjellig fra 2016 og de andre årene. Gjennomfør en test for å se om det er forskjell i inntekt mellom dagene, og mellom årene 2017, 2018 og 2019. (Hint: bruk funksjonen *linearHypothesis* fra pakken *car*).

Oppgave 4

Ledelsen i Buktafestivalen er bekymret for at det dårlige været i Tromsø påvirker salget av drikkevarer. De ønsker at du gjennomfører en test av dette. Gjennomfør en lineær regresjon hvor du har aggregert total inntekt per 15 min. I tillegg til vær legger vi til antall gjester, tid, dag og år for å passe på at forskjell som kun skyldes deltakere per år ikke fanges opp i vær-variablene. Gjennomfør regresjonen under.

$$Totalinntekt = \beta_0 + \beta_1 Nedbor + \beta_2 Lufttemperatur + \beta_3 Solskin + \beta_4 Vind + \beta_5 Gjester + \beta_6 Dag + \beta_7 \text{År}$$

Hva kan du rapporterer til Buktafestivalens styre? Hvor sikker er du på disse resultatene?

Oppgave 5

For hver festival går et stort band på scenen på lørdag klokken 21:00. Dette gjør at pilssalget klokken 20:00-21:00 er det høyeste under hele festivalen. For å forbedrede seg til pils-rushet ber Buktafestivalens styreleder deg om å predikere hvor mye pils (i antall enheter) de må gjøre klart til denne timen. Siden dette er observasjoner over et tidsintervall kan vi bruke *Poisson* fordelingen. Bruk gjennomsnittlig antall solgte pils mellom klokken 20:00 - 21:00. Hvor mange pils må gjøres klart slik at du er 95% sikker på at det ikke blir bestilt mer enn dette.