Kunskapskontroll 1 i statistik och dataanalys

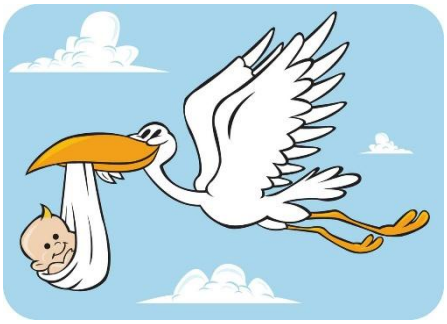
Denna kunskapskontroll består av två delar. Den första delen är teoretiska frågor kopplat till statistik. Den andra delen är kopplad till ett dataset som du kommer analysera med hjälp av Python.

Hela kunskapskontrollen lämnas in i en jupyter notebook. Betygen IG/G/VG kan erhållas. Se kursplanen för betygskriterier.

I slutet av din jupyter notebook ska du besvara följande självutvärdering:

1. Vad har varit roligast i kunskapskontrollen?
2. Vilket betyg anser du att du ska ha och varför?
3. Vad har varit mest utmanande i arbetet och hur har du hanterat det?

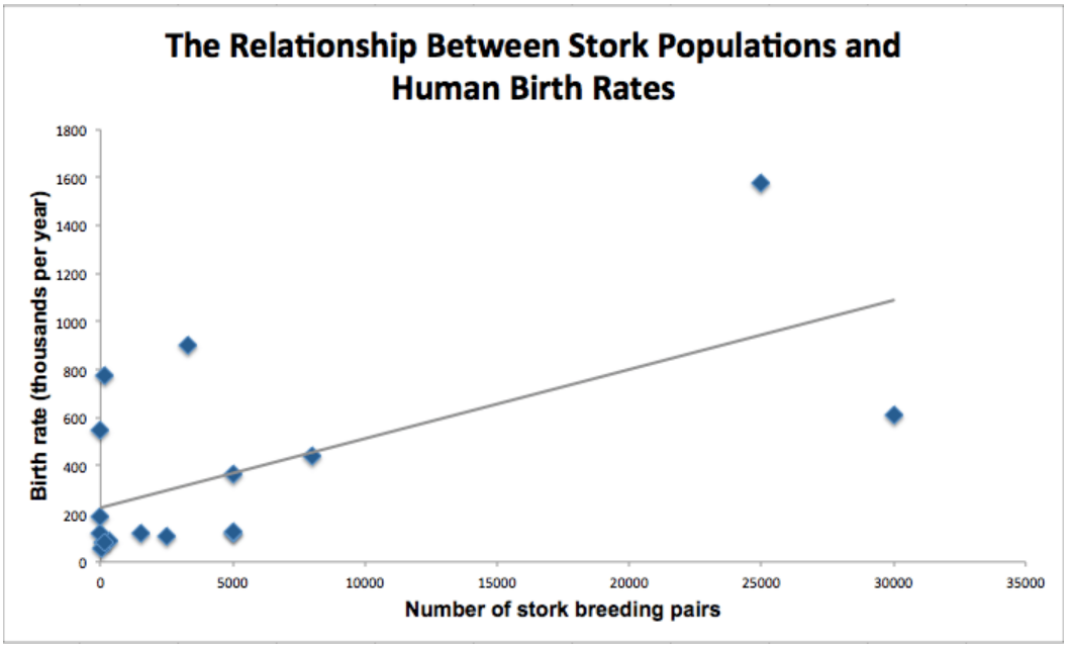
Del 1 – Teoretiska frågor



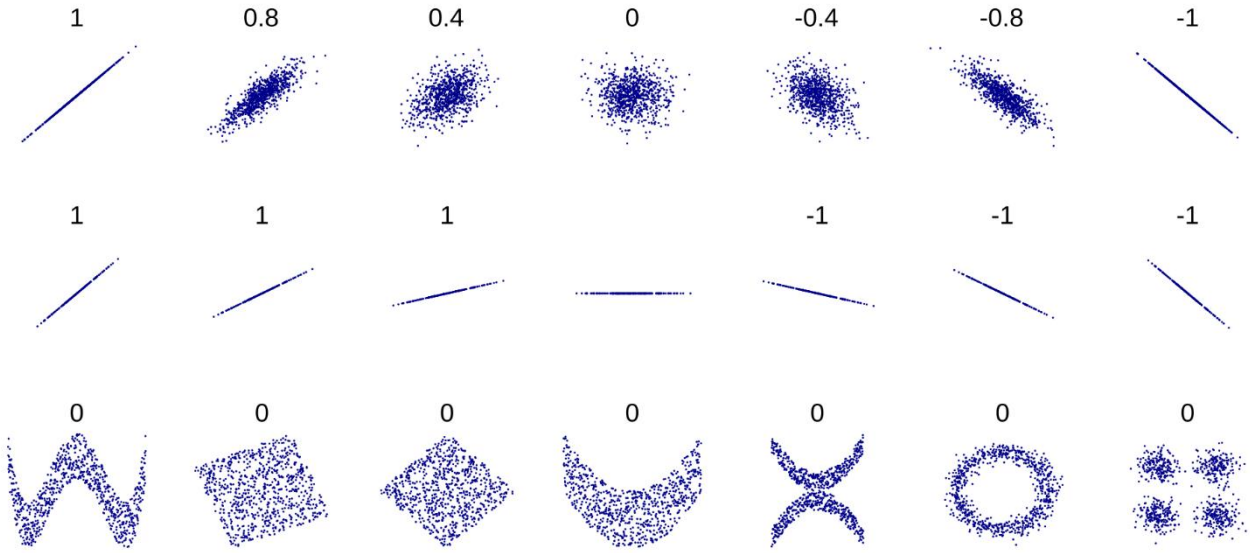
1. Många barn tror att de blev till genom att en stork kom och lämnade dem till deras föräldrar.

   För att undersöka det kan vi kolla på data och det finns en graf enligt figuren nedan.

   Det verkar alltså finnas ett samband mellan antalet storkar och barnafödsel!? Stämmer detta påståendet? Använd begreppen ”korrelation” och ”kausalitet” i ditt svar.

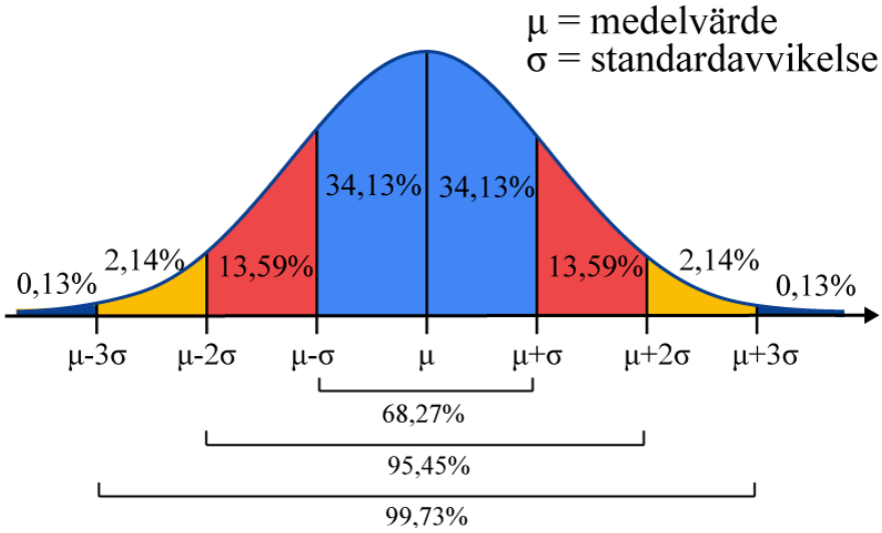


2. Din kollega som inte studerat statistik frågar dig: vad är korrelation? Förklara det för din kollega med hjälp av bilden nedan.



3. Kim påstår att ”medelvärdet” är ett bättre mått än ”medianvärdet”. Håller du med Kalle?
4. Vad används cirkeldiagram för? Ge ett exempel på vad ett företag som Spotify (eller något annat företag som du själv väljer) hade kunnat använda cirkeldiagram för.
5. Vad används linjediagram för? Ge ett exempel på vad ett företag som Spotify (eller något annat företag som du själv väljer) hade kunnat använda linjediagram för.
6. Vad används lådagram för?
7. Antag att vikten för nyfödda barn är normalfördelad. I sverige är medelvikten för nyfödda barn 3.5 kg och standardavvikelsen 0.5 kg. Vad betyder detta rent konkret? Vad är exempelvis sannolikheten att ett barn väger över 4.5 kg eller mindre än 3 kg?

   Om vi kollar på 1000 nyfödda barn, hur många barn förväntar vi oss väger över 4.5 kg?



Del 2 – Statistisk dataanalys i Python

”Data storytelling” kan definieras enligt följande: *”Data storytelling is the concept of building a compelling narrative based on complex data and analytics that help tell your story and influence and inform a particular audience.”*

- I din kod ska du först ha en kort bakgrund som förklarar vad som görs i ”jupyter notebooken”.
- Därefter gör du tillhörande dataanalys och förklarar vad som kan ses från analysen. Använd dig av beräkningar och visualiseringar.
- Slutligen presenterar du huvudinsikterna och vilka rekommendationer du ger i en ”executive summary”.

Notera, när ni gör er analys är ni fria att specificera olika antaganden eller vilka infallsvinklar ni vill ha på er analys.

De som satsar på VG behöver särskilt tänka på att lämna in en tydlig och konsistent ”Jupyter Notebook” där man kan följa med i progressionen och tankegången. Undvik alltså att göra extremt mycket beräkningar och visualiseringar utan ett tydligt syfte.

Du kan välja ett av två dataset att analysera:

1. Din vän och du är intresserade av att starta ett bilförsäljnings-företag. Innan ni gör det behöver ni förstå marknaden. Till er hjälp har ni ett dataset som ska analyseras. https://www.kaggle.com/datasets/asinow/car-price-dataset

2. VD:n på företaget ”Nytto Maximerarna” säger ofta att personalen är företagets viktigaste resurs. Och denne ber därför dig, som kan statistik och dataanalys, analysera företagets personaldata. Vilka slutsatser samt rekommendationer kan du förmedla? https://www.kaggle.com/datasets/kmldas/hr-employee-data-descriptive-analytics?resource=download