

**Mål:** Målet med uppgiften är att skriva en rapport där du svarar på en eller flera statistiska frågeställningar med hjälp av metoder vi gått igenom i kursen.

**Syfte:** Syftet med uppgiften är att prova på att arbeta med statistiska metoder och paket i Python, samt att öva på att utforska dataset och arbeta med öppna frågeställningar.

**Deadline:** Rapporten samt medföljande Python-fil(er) skall vara inlämnad uppladdad på ITHS-distans senast måndagen 6/2 kl 23.59. Rekommenderat format är PDF.

**Uppgift:** Uppgiften går ut på att ta fram en rapport som beskriver ett av tre valbara dataset med hjälp av de statistiska metoder vi går igenom i kursen. Rapporten skall vara konstruerad som ett skriftligt dokument (Word, Notebook eller Presentation) som inkluderar beskrivande text, kod-segment och plottar. Beräkningar, plottar och liknande skall göras i Python.

**Begränsningar:** Rapportens maxlängd bör vara begränsad till 5-10 sidor, beroende på format.

**Krav på rapport:** Rapporten kan utformas på valfritt sätt, men skall baseras på element från kursen, t.ex.:

- Deskriptiva mått som medelvärde, median och standardavvikelse
- Konfidensintervall
- Hypotestest
- Korrelationsanalys
- Linjär regression

**OBS! -** Rapporten behöver *inte* beskriva hela data-setet i detalj. Välj ut ett fåtal variabler/features att fokusera på. Det är viktigare med en bra frågeställning och en tydlig analys kring frågeställningen, än att inkludera ”så mycket som möjligt”.

Basera rapporten på statistiska mått och metoder vi gått igenom i kursen. Att inkludera massor av andra beräkningspaket/inferensmetoder/etc ger *inte* högre möjlighet till VG.

**Dataset:** Välj *ett* av nedanstående dataset att använda för analysen. Allihop finns tillgängliga på ITHS-distans. *Iris* och *MT-Cars* är något mindre data-set med tydliga trender, medan *Diamonds* är mycket mer omfattande och otydligt. **OBS:** Val av dataset påverkar inte bedömningen av resultatet. För de flesta studenter rekommenderas *Iris* eller *MT-Cars*, för den som vill ha mer utmaning kan *Diamonds* användas.

- *Iris* – Ett dataset som beskriver längd och bredd på blombladen hos irisblommor. Insamlat av biologen Ronald Fisher 1936.
- *MT Cars* – Data hämtad från det amerikanska magasinset *Motor Trend*. Beskriver olika motorparametrar hos 32 bilmodeller år 1973-74.
- *Diamonds* – Data om diamanter som beskriver olika fysiska parametrar så som storlek, klarhet och prissättning.

**Betygskriterier:** Inlämningsuppgiften kommer betygsättas enl. godkänd (G), väl godkänd (VG) eller icke godkänd med retur. För att bli godkänd på kursen krävs godkänd rapport. Om

rapportbetyget är VG erhålls 10 bonuspoäng till tentan. (Tentan kommer vara på totalt 50p, där gränsen för G är 25p och gränsen för VG är 37p)

För godkänd rapport krävs att rapporten innehåller beskrivande text och figurer samt fungerande, kommenterad Python-kod; att rapporten beskriver en eller flera relationer i datasetet med statistiska mått och figurer; och att minst ett konfidensintervall och/eller hypotestest utförs.

För väl godkänd rapport krävs utöver kravet för godkänd, att koden är väl kommenterad och lättläst; att figurerna är välgjorda med tydliga axlar, legender och färgsättningar; att rapporten innehåller en linjär regressionsmodell, och att modellen används till prediktion; samt att rapporten tydligt förklarar valet av mått och test som används för att besvara frågeställningen.

Om rapporten inte når upp till nivån för godkänd lämnas retur. En ny rapport skall då vara inlämnad inom en vecka efter tentadatum, varefter betyget VG inte längre går att erhålla.

**OBS!** Bonuspoängen räknas inte till nivån Godkänd på tentan, utan bidrar enbart till att nå betygsnivån Väl Godkänd.