Analys av datasetet winemag-data_first150k.csv

Om datasetet

Datasetet består av 10 kolumner samt 150930 rader. Formatet är csv. Kolumnerna är; county, description, designation, points, price, province, region_1, region_2, variety, winery.

Filen gjordes om till formatet parquet (se notebook to_parquet).

Från parquet-filen läses country, description, designation, province, region_1, region_2, variety och winery in som object. Points är dtype int64 och price är dtype float64.

Country saknar 5 värden, designation saknar 45735 värden, price saknar 13695 värden, province saknar 5 värden, region_1 saknar 25060 värden, region_2 saknar 89977 värden. Description, variety, points och winery har alla samtliga värden.

Vi har gjort om dtypes på country till category, description till string, variety till category och winery till category.

För de analyser som främst tittar på land och poäng (wine1 &2) har vi valt att ta bort kolumnerna designation, region_1, region_2, price och province pga antingen många saknade värden eller att detta inte är kolumner vi kommer gå in närmare på i dessa analyser. Kolumner som finns kvar för dessa analyser är alltså; country, description, points, variety och winery.

Vi har tagit bort de rader som saknar värden för country.

Vi har rensat dubbletter då det är ett stort antal dubbletter i detta dataset och då kvarstår 97842 rader.

För de analyser som tittar på price, points, variety och winery (wine3, 4 & 5) har vi valt att ta bort kolumnerna destination, region_1, region_2 och province då dessa inte är något vi kommer gå in djupare på i den här analysen. Kolumner som finns kvar är; country, description, points, price, variety och winery.

Vi tar bort de rader som inte har något värde för country. Samt att vi rensar datasetet på dubletter.

Därefter rensar vi bort de rader som inte har något värde för pris. Kvar finns 89125 rader. Samtliga kolumner har enbart non-null värden.

All analys i det här dokumentet görs av viner med utgångspunkt i det wine-datasetet vi har analyserat. Detta behöver alltså inte stämma överens med verkligheten då det är en samling av recensioner från en källa. Vad vi kan utläsa kring datasetet kommer dessa recensioner från amerikanska vinmagasin vilket troligen kan komma att påverka resultatet. Alla frågeställningar vi har utgår från den datan vi har i det här setet.

Notebook wine1
Analyser främst utifrån länder och produktion

Förklaring till diagrammen;
Diagram 1; Visar hur många viner som görs i de 10 länder som har störst produktion, totalt i det här datasetet finns recensioner på viner från 48 länder.

Diagram 2; Visar hur stor del av all vinproduktion de 10 största producenterna står för.

Diagram 3; Visar de 10 viner som är vanligast i de 10 länder som producerar mest, procentuell jämförelse med övriga viner.

Diagram 4; Visar hur stor del av den totala mängden vin som de 5 vanligaste utgör.

Diagram 5; Visar hur stor del av den totala mängden vin som de 10 vanligast utgör.

Diagram 6; Visar vilka 5 länder som har flest viner med 99 eller 100 poäng.

Analys med frågeställningar;
1. Hur många viner görs det i de länder som har störst vinproduktion?
Diagram 1 och 2; Vi ser att USA står för en klar majoritet av vintillverkningen, drygt 40 000 av de nästan 98 000 vinerna som finns med i det här datasetet. USA producerar mer än dubbelt så mycket som de som kommer på 2a, 3e, 4e och 5e plats tillsammans. De 10 största vinproducerar-länderna står för 95% av all vintillverkning. Vad vi inte kan veta utifrån den här analysen är hur väl detta stämmer överens med den faktiska vintillverkningen i världen. Dock kan man tänka sig att det kan finnas en viss bias då USA står för en väldigt stor del av det här datasetet. Detta är en osäkerhet i datasetet.

2. Vilka viner producerar de länder som har störst vinproduktion?
Diagram 3; Här ser vi en större spridning över vilka sorter som produceras även om de som ligger på top 10 bland de vanligaste i de länder med störst produktion utgör en stor del av produktionen. 66.1% av allt vin dessa länder producerar hör till de 10 vanligaste.
Inte särskilt överraskande ser vi även, när vi tittar på hela mängden viner för hela datasetet, att de 10 vanligast förekommande är desamma som är de 10 vanligast förekommande hos de 10 största vinproducentländerna. Detta är förväntat i och med att de 10 länder med störst vinproduktion står för en så betydande del av produktionen av vin.

3. Är det någon skillnad på vilka viner som produceras beroende på om vi tittar på hela datasetet eller tittar på de 10 länderna med störst produktion?
Diagram 4 och 5; Vi ser en minimal skillnad mot tidigare diagram när vi tittar på hela datasetet, avseende de 5 samt 10 vanligaste vinerna. Det som är intressant är även att ha med sig detta i kommande analyser då de vanligast förekommande vinerna troligen kommer ha en stor spridning i både betyg och pris eftersom det finns många varianter av dessa, medan de mindre vanligt förekommande troligen kan ha en mindre spridning då varianterna är mindre vanligt förekommande och vissa sorter enbart finns i en variant.

4. Vilka länder har flest viner med höga poäng?
Diagram 6; visar tydligt att de med flest viner med höga poäng kommer från USA, Italien och Frankrike. Detta är inte särskilt förvånande då de har stor produktion och därmed har många sorter med många varianter. Det man kan tänka är att variationen på poäng även bör vara stor och därmed bör dessa även vara högst representerade med höga poäng. Dock är det tydligt att enbart storlek på produktion inte är det som påverkar mest då USA har en så pass mycket större produktion än både Italien och Frankrike men ändå inte sticker ut på samma sätt när man tittar på antal viner med de högsta poängen. Dvs att tittar man procentuellt på antal viner med höga poäng motsvarar inte det den procentuella tillverkningen sett utifrån land.

Notebook wine2
Analyser främst med utgångspunkt i poäng

Förklaring till diagrammen;
Diagram 1; Visar fördelning av poäng hos de vinsorter som förekommer med 100 poäng.

Diagram 2; Visar medelvärde på de viner som förekommer med 100 poäng.

Diagram 3; Viner med högst medelbetyg

Övriga kodblock; visar förekomsten av viner med högst poäng och om de även förekommer bland de 10 vanligaste vinerna. Samt om de sorter som förekommer med högst poäng även återfinns i de med lägst poäng.

Diagram 4; Medelpoäng för de vanligast förekommande vinerna.

Analys med frågeställningar;
1. Vilka viner förekommer med 100 poäng och hur ser fördelningen av poäng ut för dessa sorter?
Diagram 1; Det man kan se här är att de med många sorter har en tydligare normalfördelning vilket var väntat i och med att de finns större mängd data att analysera. Samtliga som har minst en variant med 100 poäng förekommer även med betydligt lägre poäng vilket gör det tydligt att det är viktigt att veta vilken variant av ett visst vin man ska välja för att få den bästa upplevelsen. Det man också kan se är att majoriteten av dessa viner har större delen av poäng under 90 poäng vilket förstärker vikten av att veta vilken variant av vinet man väljer.

2. Har vinsorterna som förekommer med höga poäng även höga medelpoäng?
Diagram 2; Ser man här på medelvärdena av vinerna som förekommer med högsta poäng i någon variant så ser man att medelvärdet ligger betydligt lägre och endast två av sorterna har ett medelvärde över 89 poäng, vilket är det högsta medelvärdet som förekommer för dessa viner. Även detta förstärker insikten att det är mer intressant att utgå från variant av vinsort för att hitta viner med högsta poäng än att enbart gå efter vilken sort som förekommer med höga poäng.

3. Har de som förekommer med högsta poäng högra medelvärde än andra viner i datasetet?
Diagram 3; Det här diagrammet visar top 10 vinerna sett till medelbetyg. Ingen av dessa finns med bland de vanligast förekommande och inte heller bland de med någon variant som har 100 poäng. Det man kan tänka sig är en anledning till detta är att det är viner med en mindre produktion där en eller flera varianter av vinerna har höga poäng vilket ger ett högt medelvärde. För att kunna dra slutsatser av detta skulle man behöva gå in djupare i hur stor produktion dessa viner har och om det enbart är en eller ett fåtal enstaka varianter eller om det finns sorter som har en relativt hög variation av varianter som återfinns bland de med högst medelvärde.

4. Finns de sorter med 100 poäng även med bland de sorter som är vanligast förekommande samt finns de med bland de sorter som även förekommer med lägst poäng?
Det man också ser är att de med 100 poäng även i relativt hög grad förekommer bland de som är mest vanliga. 6 sorter av de som förekommer med 100 poäng finns med bland de 10 vanligaste. Detta i sig är inte förvånande då man kan tänka att de med hög produktion har en stor spridning både uppåt, med höga poäng och nedåt med lägre poäng.
Vi ser också att 8 av de sorter som förekommer med högst poäng har varianter som återfinns bland de med lägst poäng. Återigen visar det på vikten av att veta vilket vin man väljer för att få en så bra upplevelse som möjligt. Även detta var ganska väntat i och med att flera av de viner som återfinns med varianter med 100 poäng har en relativt stor produktion och därmed också en trolig spridning på poäng.

5. Hur ser medelpoängen ut för de vanligast förekommande vinerna?

Det man kan se här är att de vanligast förekommande vinerna har en relativt stor spridning. Flera av dessa som vi konstaterat tidigare förekommer med riktigt höga poäng men även med låga vilket gör att det blir intressant att se medelvärdena för dessa. Det man kan konstatera är att de verkar ha en stor spridning i sina poäng med både låga och höga poäng vilket gör att medelvärdena hamnar från strax över 86 upp till medelpoäng strax över 89. Detta är relativt förväntat eftersom de har många varianter och därmed kommer det högst troligt finnas en stor spridning på poäng.

Notebook wine3
Analyser med utgångspunkt i poäng och pris

Förklaring till diagrammen;

Diagram 1; Visar jämförelse mellan vanligt förekommande viner och mindre förekommande viner, pris och poäng, Antal sorter; 1-10, 10-100, 100-500, 500-1000 och över 1000

Diagram 2; Visar Medianpris och medelpris på viner, grupperat per land.

Diagram 3; Visar medelpris, medianpris, min och max för viner uppdelat på poäng.

Diagram 4; Visar vinerna, grupperat på sort med högst medelbetyg. Visar betyg och pris. Vinsorter som förekommer med minst 5 varianter

Diagram 5; Visar vinerna, grupperat på sort med lägst medelbetyg. Visar betyg och pris. Vinsorter som förekommer med minst 5 varianter

Diagram 6; Visar vinerna, grupperat på sort med högst medelbetyg. Visar betyg och pris. Alla viner.

Diagram 7; Visar vinerna, grupperat på sort med lägst medelbetyg. Visar betyg och pris. Alla viner.

Diagram 8-17 ; Visar de vanligast förekommande vinerna och distributionen av poäng och pris för dessa.

Diagram 18; Visar viner som är dyrare än 500 USD, grupperat per land.

Diagram 19; Visar medelbetygen för de 5 vanligaste vinerna, grupperat på land

Diagram 20; Visar medianpriset för de 5 vanligaste vinerna, grupperat på land

Diagram 21- 30; Visar de 5 vanligaste sorterna och medelbetyg samt medianpris, uppdelat per vinsort.

Analys med frågeställningar;

1. Finns det stora skillnader mellan vinerna baserat på hur många varianter det finns av sorterna?
Diagram 1; Det man kan se i det här diagrammet är att poängen är relativt lika fördelade oavsett hur många varianter av sorter som finns av vinerna. Samtliga har minst 1 med höga poäng, minst 1 med låga poäng och medel samt medianpoängen ligger relativt lika oavsett hur många varianter det finns av vinerna. Däremot så skiljer sig maxpriserna stort. Ju större variation av sorterna det finns desto dyrare viner finns det. Detta kan bero på att de med färre varianter ev produceras hos mindre vinproducenter där det kanske är svårare att ta ut samma priser som de större och mer välkända. De som har fler varianter av sina vinsorter kanske också specialiserat sig på just detta och har flera varianter av samma sort, med olika lagring och olika processer vilket gör att de kan ha ett större spann på priserna. Att en vinsort är vanligare skulle också kunna vara en fördel för att kunna ta ett högre pris då fler känner igen och vet vad de kan förvänta sig av just det vinet.

2. Är det stora skillnader på medianpris och medelpris på viner beroende på vilket land de produceras i?
Diagram 2; Vad man kan se här är att det skiljer sig stort mellan länderna som förekommer i det här datasetet. Man ser också att medelpriset, så gott som i alla länder, är högre än medianpriset. Orsaken är troligen att länderna har några viner som är betydligt dyrare än andra vilket gör att medelvärdet ökar. Troligtvis ger medianpriset en mer korrekt bild här för att få den mest korrekta

bilden av vad ett vanligare vin kan kosta i respektive land. Den här jämförelsen är enbart baserad på pris och säger inget om poängen eller vilken sort av vin det gäller och visar därmed enbart variationen av pris för ländernas samtliga viner. Länderna har också olika stor produktion av vin och detta kan även påverka spridningen i pris. Priserna skiljer sig en hel del mellan olika länder vilket är intressant för att göra vidare analyser för att titta närmre på vilka viner som produceras i vilka länder (detta har vi dock inte utforskat i det här arbetet).

3. Hur ser skillnaderna ut i medelpris, medianpris, maxpris och minpris baserat på poäng de olika vinerna har?
Diagram 3; Det man kan se i det här diagrammet är att det är en tydlig ökning i medelpris samt medianpris för vinet allt eftersom man kommer upp på högre poäng. En annan aspekt som är intressant i det här diagrammet är att titta mer på min och maxpris. Det man kan se är att det finns viner med låga poäng som kostar betydligt mer än vissa med höga poäng vilket visar att man kan hitta riktigt bra vin även till billigare pris samt att man kan hitta de med låga poäng som är riktigt dyra och därmed eventuellt mindre prisvärda. Spridningen i pris på viner med samma poäng är otroligt stor. Det skulle kunna bero på vilken sort det är och att vissa välkända viner går att ta bra betalt för även om de inte är de bästa utan mer baserat på att de är kända, samtidigt som andra riktigt bra viner inte är lika kända och därav inte möjligt att ta samma pris för. Det man också behöver ha med sig är att vi i den här analysen helt går på vinrecensenternas bedömning och den är trots allt subjektiv. Detta skulle också kunna vara en del i förklaringen då bedömningen inte nödvändigtvis skulle vara likadan för andra vinrecensenter.

4. Har de vinsorter med högst medelpoäng även höga medelpriser?
Diagram 4+6; Det man ser är att vinerna med högst medelpoäng har relativt lika medelpoäng medan medelpriset diffar mycket. Det som också är intressant är att dessa inte förekommer i vare sig de vanligaste eller de som har minst en variant med 100 poäng. Ändå hamnar de högst i medelpoäng. Orsaken skulle kunna vara att det är mindre sorter vilket gör att spridningen inte blir lika stor. I diagram 6 där alla vinsorter är med oberoende på antal varianter är att en sort sticker ut, Cabernet-Shiraz då medelpriset för denna är betydligt högre än för de andra. Dock återigen så kan detta bero på att det är en känd producent samt ett vin med höga poäng vilket gör att priset kan sättas därefter. Vi tittar här på medelpriset och är det en enstaka variant som sticker iväg i pris så kan det också förklara varför medelpriset är högre.

5. Har de vinsorter med lägst medelpoäng även låga medelpriser?
Diagram 5+7; Vinerna med lägst medelpoäng har generellt en betydligt lägre kostnad vilket inte är förvånande då lägre betyg ofta kan innebära billigare vin då man inte är lika benägen att betala mycket för något som inte är en lika bra upplevelse. Dessa viner är heller inga av de vanligast förekommande, utan troligen mindre varianter som kanske inte har samma spridning på poäng då det helt enkelt inte produceras i samma mängder.

6. Har de vanligaste vinsorterna en liknande fördelning av poäng samt pris?
Diagram 8-17; De här diagrammen visar en ganska tydlig normalfördelning över poängen, om än lite förskjuten mot de nedre poängen. Detta skulle kunna förklaras av att det är en stor tillverkning av dessa vinsorter samtidigt som det är svårt att producera vinerna som når de högre poängen och därmed finns en stor produktion där man hamnar på aningen lägre poäng samtidigt som man bland alla dessa varianter även lyckas producera varianter med högre poäng.
Det man kan se är att priserna skiljer sig relativt mycket, framförallt maxpriserna är höga på några sorter, ex Chardonnay och Bordeaux-style Red Blend förekommer med priser över 2000USD, medan Zinfandel har ett maxpris på dryga 100. Däremot är spridningen i priser stor även om de flesta viner har en tydlig majoritet av varianter som ligger i en lägre prisnivå (under 50USD samt under 100 USD).

Man kan tydligt se hur majoriteten av varianterna ligger i de lägre prisspannen även om det utifrån de här diagrammen är svårt att utläsa hur låga priserna är på några av vinerna då spannet sträcker sig från 0-2500. För att få en tydligare bild av priserna och hur de är fördelade på de sorter av viner som förekommer med höga priser skulle man behöva göra en ytterligare analys för att se spridningen över de lägre spannen av priser.

7. Vilka viner har priser över 500 USD och hur är de fördelade per land? Dvs vilka länder har flest riktigt dyra viner?

Diagram 18; Det är i det här diagrammet tydligt att Frankrike har flest dyra viner med närmare 30 st varianter över 500 USD. Det man kan tänka sig är att Frankrike är ett känt vinland med sorter som är dyrare än på andra ställen. Möjligen kan det bero på produktion av viner och att det är anrika vingårdar där man har mycket kända viner, vilket i sig kan göra att man kan ta höga priser för dessa. Övriga länder har betydligt färre viner som kostar över 500USD, samtidigt är majoriteten av länderna kända vinnationer samt majoriteten även länder där produktionen är stor. Något som också är intressant utifrån diagram i inlam_wine1 är att det är USA som har flest viner med 99 eller 100 poäng. Dock är det Frankrike som har dyrast viner. Detta är lite förvånande men den slutsats man kan dra är att poängen nödvändigtvis inte alltid behöver betyda att vinerna är jättedyra.

8. Hur ser medelbetyg och medianpris ut för olika länder när det kommer till de 5 vanligaste vinsorterna?
Diagram 19-20; Det som är intressant med dessa diagram är att de länder som har högst medelpoäng inte också har högst medianpris. Det är relativt stora skillnader och de som sticker ut i medelpoäng gör det inte i medianpris utan där är det andra länder som har betydligt högre priser. Det är intressant att se den här variationen då man skulle kunna tänka sig att de med höga medelpoäng även skulle ha ett högre medianpris. De två med riktigt höga medianpriser och de som sticker ut mest är Canada och Turkiet vilket inte är länder med en stor produktion av vin. Det man kan tänka här är att det kan vara få varianter av sorter som är med i datasetet och om någon/ några av dessa sticker iväg i pris kan det göra att medianpriset blir högt. De tre länder med störst produktion ligger ganska lika i medianvärde vilket är förväntat med tanke på stor produktion och att det därmed bör finnas många varianter av vinerna och därmed ett stort spann av priser.

9. Kan vi se liknande mönster avseende medelbetyg och medianpris om man tittar på de 5 vanligaste vinsorterna för sig?
Diagram 21-30; Det vi kan se här är att alla länder inte producerar alla viner. Rent generellt kan man inte se något mönster att något land alltid har högst medelpoäng eller medianpris utan det verkar vara mycket beroende på sort. Vill man gå in djupare i en analys för detta så är troligen det bästa att titta på länderna var för sig för de länder man är intresserad av och då ställa alla vinsorter landet producerar mot varandra. De här större diagrammen ger en bra överblick över att det skiljer sig från vinsort till vinsort.

Notebook wine4
Analyser med utgångspunkt i Vingårdar

Förklaring till diagrammen i wine4;

Diagram 1-11; Visar vilka viner respektive vingård har. Vingårdar med minst 1 variant som har fått högsta poäng.
Övrigt cellblock; Frågeställningen; hur många sorter producerar de största vingårdarna? Och vilka vingårdar har störst produktion?

Förklaring till diagrammen i wine_spark;

Diagram 1; Visar de 10 vingårdarna med högsta betyg.

Diagram 2-11; Visar de 10 största vingårdarna med vilket vin de tillverkar samt poäng.

Analyser med frågeställningar;

1. Finns det någon gemensam faktor utifrån antal sorter, fördelning av poäng eller pris för de vingårdar som har en variant av vinsort som förekommer med 100 poäng? inlam_wine4
Det som är intressant här är att se hur det skiljer sig väldigt mycket mellan de olika vingårdarna. Några har stor produktion av många sorter medan andra har ett fåtal sorter. Det som är intressant med det är att man kan tro att vissa specialiserar sig på den sorten som är på topp och därmed har en mindre produktion. Andra har stora variationer i både sorter och antal där den med högst poäng endast är en liten del. Det som också är intressant med de som har fler sorter är att poängen och priserna på dessa skiljer sig mycket och det finns inget tydligt som är gemensamt för alla vingårdar som har en variant som fått 100 poäng. Vi har i den här analysen tagit bort de viner som saknar priser vilket gör att vissa av vingårdarna har sorter som har fallit bort i den här filtreringen. Dock ville vi här jämföra både poäng och priser för att få en bild över båda dessa parametrar. För ytterligare analyser kan det dock vara intressant att gå in mer på djupet för sorter oberoende av pris.

Det som också är en intressant aspekt är att det enbart är en av vingårdarna som är bland de största vingårdarna som har ett vin med 100 poäng. Övriga vingårdar som har ett vin med toppbetyg finns därmed inte bland de 10 största producenterna.

De 10 vingårdar som tillverkar flest viner producerar totalt 2379 varianter, vilket är en liten del av alla vinerna i datasetet. Av detta kan vi dra slutsatsen att en enskild vingård inte bör påverkar datasetet i hög grad.

2. Vilka vingårdar har högsta medelbetyget på sina viner? wine_spark
Det som är intressant att se här är att enbart en av de vingårdar som förekommer med en vinsort med 100 poäng återfinns bland de med högst medelpoäng. Detta är även en vingård som enbart har en sort vilket självklart påverkar poängen. Därmed är det svårt att säga hur stor vikt man ska lägga vid detta. En slutsats är snarare att de vingårdar med högst medelpoäng kan tänkas ha flera sorter med höga poäng om än inte 100.

3. Har de största vingårdarna fler varianter av sorter med höga poäng? wine_spark
Det som är intressant att se här är att det är flera av dessa vingårdar som har många sorter och även en stor spridning av poäng på sina varianter av samma sort. Både höga och lägre poäng. Detta är relativt förväntat i och med att de med så stor produktion med all sannolikhet har en stor variation på lagring, druvor, tillverkningssätt. Det man också kan tänka är att de med högre poäng ev förekommer i en mindre produktion då de troligen är dyrare att tillverka. Som en bas i tillverkningen kan man då tro att de har en större kvantitet av flera olika sorters viner.

Notebook wine5

Analyser med utgångspunkt i korrelationer

Förklaring till diagrammen;

Diagram1; Visar korrelation mellan pris och poäng på de 100 dyraste vinerna, grupperat på sort.

Diagram 2; Visar korrelation på de 100 billigaste vinerna mellan pris och poäng, grupperat på sort.

Diagram 3; Korrelation på de viner som förekommer med 100 poäng, mellan pris och poäng.

Analyser med frågeställningar;

1. Finns det någon korrelation mellan pris och poäng för de 100 dyraste vinerna, grupperat per sort?
Diagram 1; Tittar man på de 100 dyraste vinerna, grupperat på sort ser man en stark korrelation för vissa sorter, dock har majoriteten inte någon eller en negativ korrelation mellan pris och poäng. Slutsatsen är att det är svårt att säga att det är en tydlig korrelation mellan pris och poäng för de 100 dyraste sorterna generellt, däremot på en nivå där man tittar på sort för sort kan man se ett samband. Man kan alltså säga att det generellt inte finns någon korrelation mellan pris och poäng för de 100 dyraste viner.

2. Finns det någon korrelation mellan pris och poäng för de 100 billigaste vinerna, grupperat per sort?
Diagram 2; Samtliga visar ingen eller en negativ korrelation mellan pris och sort. Detta gör att man kan dra slutsatsen att man inte kan se någon korrelation mellan pris och poäng för de 100 billigaste vinerna generellt. I så fall snarare en negativ korrelation vilket är intressant då man annars skulle kunna tänka att låga priser skulle innebära låga poäng.

3. Finns det någon korrelation mellan pris och poäng hos de viner som förekommer med 100 poäng på minst 1 variant?
Diagram 3; Det finns en viss korrelation mellan pris och poäng för de här vinerna. Prugnolo Gentile sticker ut med en korrelation på 0,79. Övriga ligger på en korrelation mellan 0.42 och 0.61 vilket ändå kan ses som en viss korrelation mellan pris och poäng.