1. Teoretiska frågor

1. Data lagras med vissa relaterade attribut, ofta något slags ID-nummer, på så sätt kan man lätt hitta annan data som är relaterad till det du söker, därav “relations”-databas
2. CRUD är ett akronym för följande; C - Create, R - Read, U - Update, D - Delete. Beskriver generellt vad man kan göra i databasen.
3. Joins för samman två tabeller, left-join tar med alla specificerade attribut i “vänstra” tabellen men bara de identiska attribut som finns i den högra, användbart när du vill ta fram en hel tabell och se hur den korrelerar med en annan. Inner-join tar med de attribut som är identiska i båda tabellerna, användbart när du vet exakt vilka attribut du vill använda.
4. Indexering i SQL används för att snabba upp query-processen, man tilldelar värden till data och på så sätt behöver inte SQL leta igenom hela tabeller eller databaser utan vet vart det specifika indexet finns. För och nackdelar finns, men indexering är mest användbart när man arbetar med stora databaser.
5. En vy är som en “sparad”, förinställd query som är redo att skickas till databasen, användbart om man behöver titta på samma data väldigt ofta.
6. En lagrad procedur i SQL är “lagrad kod” vilket är användbart om man kör samma kod ofta, då kan man spara koden i en lagrad procedur vilket gör den lätt att använda för andra användare. Lagrade procedurer liknar funktioner, men det finns subtila skillnader.

2. Programmeringsuppgift och rapport

1. Deskriptiv sammanfattning

Syftet med den här rapporten är att ge en överblick över företaget AdventureWorks, ett företag verksamt inom försäljning av cyklar och diverse cykelrelaterade tillbehör. Rapporten innehåller dessutom några statistiska analyser av den data som finns tillgänglig, samt rekommendationer baserade på dessa statistiska analyser.
AdventureWorks databas innehåller mycket information såsom omfattande information om dess kunder, produkter, produktkategorier och information om enskilda ordrar kunderna lagt. I databasen finns även information om de anställda på AdventureWorks. Den lagrade datan är daterad till tidsspannet 2011-2014 och alla siffror angedda antas vara i US dollar då majoriteten av kunderna befinner sig i USA.

Det första jag ville göra var att fastställa vilka produkter företaget säljer samt vart de är verksamma, för att lättare kunna avgöra relaterade frågor såsom vilken valuta som anges, hur fraktsituationen/kostnader ser ut och hur företaget generellt mår, finansiellt sett.

Därefter ville jag ta reda på vilka produkter som säljer bäst, med fokus på vinstgenerering, samt titta på om det finns produkter som är lönsamma men inte säljer bra och på så sätt eventuellt skulle kunna öppna för nya segment att inrikta sig på.

Jag ville även titta på hur stort produktbortfallet var och om man kunde försöka prediktera detta bortfall framåt i tiden för att därefter kunna anpassa produktion/inköp och få ner dessa siffror.

2. Statistisk analys

	Name
0	Bib-Shorts
1	Bike Racks
2	Bike Stands
3	Bottles and Cages
4	Bottom Brackets
5	Brakes
6	Caps
7	Chains
8	Cleaners
9	Cranksets
10	Derailleurs
11	Fenders
12	Forks
13	Gloves
14	Handlebars

Här visas ett urval av den första frågan som ställdes mot databasen, då jag ville veta inom vilka kategorier företagets produkter finns.
Vi ser tydligt att företaget enbart är verksamma inom försäljning av cyklar och cykelrelaterade tillbehör.

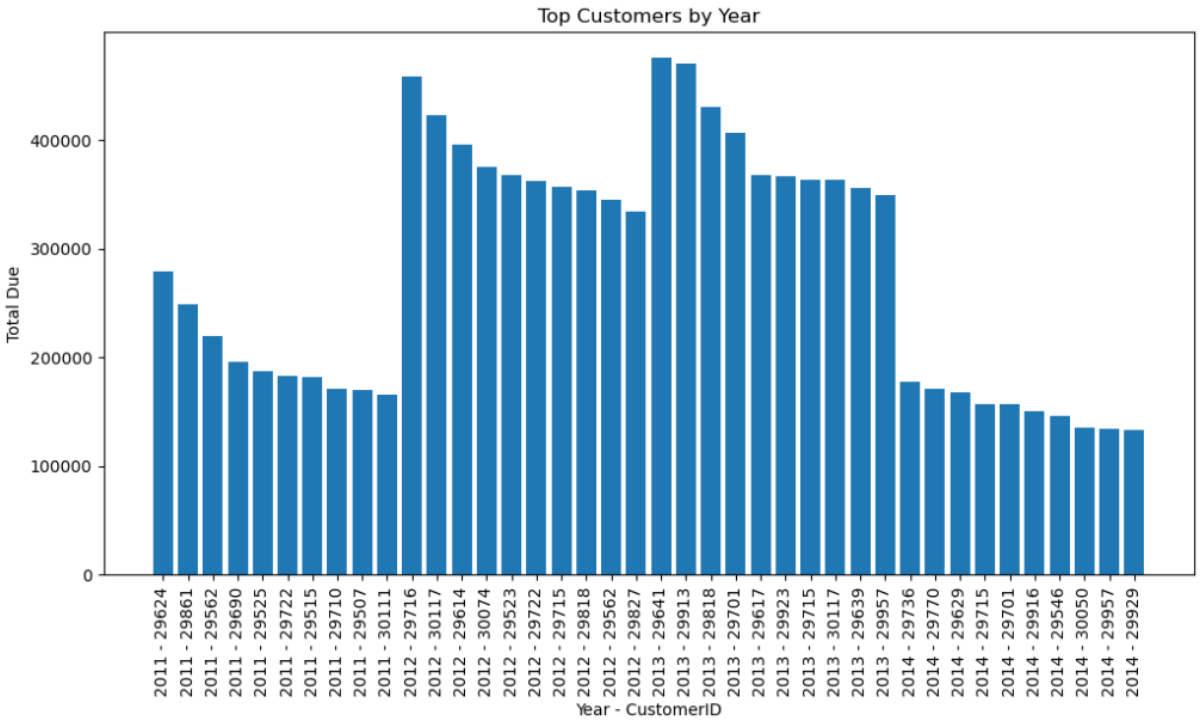
Bild 1

Därefter tittade jag på företagets toppkunder och hur mycket de handlat för under den treårsperiod som databasen innefattar.

Bild 2

	CustomerID	TotalDue	TotalOrders	AvgOrderValue
0	29818	989184.0820	12	82432.0068
1	29715	961675.8596	12	80139.6549
2	29722	954021.9235	12	79501.8269
3	30117	919801.8188	12	76650.1515
4	29614	901346.8560	12	75112.2380
5	29639	887090.4106	12	73924.2008
6	29701	841866.5522	8	105233.3190
7	29617	834475.9271	12	69539.6605
8	29994	824331.7682	12	68694.3140
9	29646	820383.5466	12	68365.2955

Bild 3



Sedan ville jag analysera om det finns trender i vilka kunder som handlar mest, men det ser ut att variera från år till år.

	SalesPersonID	SalesMade	TotalAmount	Location	Gender	Name
0	276	418	10367007	Southwest US	F	Linda Mitchell
1	277	473	10065803	Central US	F	Jillian Carson
2	275	450	9293903	Northeast US	M	Michael Blythe
3	289	348	8503338	United Kingdom GB	F	Jae Pak
4	279	429	7171012	Southeast US	M	Tsvi Reiter
5	281	242	6427005	Southwest US	M	Shu Ito
6	282	271	5926418	Canada CA	M	José Saraiva
7	290	175	4509888	France FR	M	Ranjit Varkey Chudukatil
8	283	189	3729945	Northwest US	M	David Campbell
9	278	234	3609447	Canada CA	M	Garrett Vargas
10	280	95	3325102	Northwest US	F	Pamela Ansman-Wolfe
11	284	140	2312545	Northwest US	M	Tete Mensa-Annan
12	288	130	1827066	Germany DE	F	Rachel Valdez
13	286	109	1421810	Australia AU	F	Lynn Tsoflias

Bild 4

Sedan tog jag fram info om vilka individuella säljare som sålt mest genom databasens livslängd och vilket geografiskt område de är tilldelade, här drar jag slutsatsen att AdventureWorks troligtvis är ett amerikanskt bolag som handlar i US dollar främst. Notera att ungefär en fjärdedel av den totala försäljningen inte sker via säljare, utan troligtvis via egen butik/hemsida.

Results Messages

	SoldBySalesPeople
1	90775446,9931

	NotSoldBySalesPeople
1	32441339,1228

Bild 5

	ProductID	Name	TotalUnitsSold	StandardCost	ListPrice	ProfitMarginPct	TotalRev
0	782	Mountain-200 Black, 38	2977	1252	2295	45	1305847
1	783	Mountain-200 Black, 42	2664	1252	2295	45	1227621
2	779	Mountain-200 Silver, 38	2394	1266	2320	45	1153481
3	781	Mountain-200 Silver, 46	2216	1266	2320	45	1111307
4	784	Mountain-200 Black, 46	2111	1252	2295	45	1104546
5	780	Mountain-200 Silver, 42	2234	1266	2320	45	1096545
6	753	Road-150 Red, 56	664	2171	3578	39	668314
7	749	Road-150 Red, 62	600	2171	3578	39	661279
8	794	Road-250 Black, 48	1498	1555	2443	36	632542
9	793	Road-250 Black, 44	1642	1555	2443	36	626323

Eftersom tre år är en ganska begränsad tidsperiod valde jag istället att analysera lönsamheten i bolaget, på bild 5 ser vi att de kompletta cyklarna genererar mest vinst och att de bästsäljande cyklarna har en vinstmarginal på 45%.

	ProductID	Name	StandardCost	ListPrice	ProfitMarginPct	TotalRev
0	710	Mountain Bike Socks, L	3	10	64	269
1	709	Mountain Bike Socks, M	3	10	64	1147
2	879	All-Purpose Bike Stand	59	159	63	24784
3	865	Classic Vest, M	24	64	63	22062
4	876	Hitch Rack - 4-Bike	45	120	63	59796
5	877	Bike Wash - Dissolver	3	8	63	6604
6	866	Classic Vest, L	24	64	63	7990
7	930	HL Mountain Tire	13	35	63	30586
8	864	Classic Vest, S	24	64	63	27110
9	878	Fender Set - Mountain	8	22	63	29184

Bild 6

Om vi istället sorterar efter vinstmarginal ser listan ut såhär och vi kan tydligt se att bland annat kläder och vissa tillbehör har väldigt hög vinstmarginal men låga försäljningssiffror.

	ProductID	TotalQty	TotalScrapped	Mean	STD	OrderCount	SE	CI95HI	CI95LO
0	331	236002	1374	215	1146	1093	35	280	150
1	532	469468	1154	429	2277	1093	69	494	364
2	3	911890	1031	834	4143	1093	125	899	769
3	316	236002	736	215	1146	1093	35	280	150
4	350	118001	692	107	573	1093	17	172	42
5	324	234734	585	214	1139	1093	34	279	149
6	327	117367	571	107	569	1093	17	172	42
7	529	94218	422	86	433	1093	13	151	21
8	399	117367	348	107	569	1093	17	172	42
9	533	117367	342	107	569	1093	17	172	42

Bild 7

I bild 7 ser vi ett konfidensintervall för databasens 3 år som innehåller de produkter som skrotas på grund av olika anledningar.

3. Slutsatser och rekommendationer

Min första tanke efter att jag analyserat databasen är att tidsramen är relativt kort och att det är svårt att utläsa några trender utifrån endast tre år, därför har jag snarare fokuserat på att analysera på produktnivå för att enklare se vart lönsamheten finns i bolaget.
Vidare ser vi från Bild 5 att de kompletta cyklarna genererar mest vinst i bolaget, förmodligen för att de är tillgängliga även för de som inte kan något om cyklar och att de därför tilltalar en bredare population.

Om vi däremot tittar på Bild 6 ser vi att de högsta marginalerna finns vid bland annat kläder och diverse tillbehör, men vinsten genererad från framförallt kläd-segmentet är väldigt lågt eftersom de sålts i väldigt små kvantiteter, min rekommendation är att utforska det här segmentet och se om man kan satsa mer på det, åtminstone de produkter som visats sälja dåligt. Till exempel kan man köra erbjudanden på klädsortimentet och erbjuda paketerbjudanden tillsammans med de kompletta cyklarna för att öka synligheten för det här segmentet.

Enligt Bild 2 och 3 påvisas att kundbasen är väldigt spridd, utan tydliga trender i deras inköp. Att kundbasen är så pass spridd är ett gott tecken, eftersom beroendet av en enskild kund inte alls är högt och därför minimeras risken för AdventureWorks ifall en av deras kunder får problem. Konfidensintervallet för skrotade produkter kan hjälpa identifiera vart man bör lägga extra resurser för att få ner antalet skrotade produkter, jag rekommenderar att börja med ProduktID 3, då den ser ut att ha högre “skrotningsgrad” än resterande produkter.

4. Executive summary

- Tidsramen är begränsad till 3 år, vilket gör det svårt att se långsiktiga trender.
- Kompletta cyklar genererar mest vinst, troligtvis på grund av dess “turnkey solution”.
- Trots högst marginaler i klädsegmentet är vinsten här låg, jag rekommenderar att öka marknadsföring samt erbjuda paketerbjudanden med cykelsegmentet.
- Ingen tydlig inköpstrend finns bland kundbasen, vilket indikerar spridd efterfrågan.
- Datan visar att AdventureWorks har en bred kundbas, vilket minskar riskerna som kan finnas när man är beroende av storkunder.
- Sammantaget visar analysen att AdventureWorks har goda förutsättningar att nå en stark marknadsposition då det finns potential för ytterligare tillväxt och lönsamhet inom vissa segment.
- Analysera konfidensintervallet för att hitta avvikelser, åtgärda de största avvikelserna först.

5. Datum för muntlig presentation

Muntlig presentation genomfördes 09:00-09:20 den 3/1-24.

3. Reflektion på eget arbete

1. **Utmaningar;**

- Satt ganska länge och försökte ansluta till SQL Server via SQLAlchemy, av någon anledning hittade den inte anslutningen. Tillslut gav jag upp, gjorde om alla steg från början och lyckades då. Oklart varför, koden var densamma.
- Hade problem i queryn där jag tittade på vinstmarginal och vinstgenerering. Fick ett error som jag trodde hade med summering av “TotalRevenue” att göra, så jag skapade en CTE vilket löste problemet.
- Det var inte alltid lätt att tolka infon i vissa tabeller, till exempel när man vill koppla ihop tabeller som har flera steg mellan sig, eller när kolumner har otydliga namn. Då får man försöka hitta beskrivningar eller testa sig fram.
- Det tog också lång tid för mig att förstå hur SQL-Python-Pandas-MatPlotLib interagerade med varandra, eftersom det verkar finnas många olika sätt att koda detta, därför tog det också väldigt lång tid att få ihop den bar chart som visar kunders inköp på årsbasis.
- När jag beräknade vinstmarginal och total vinst fick jag “Divide by zero” error, Mark gav mig lite pointers och tillslut hittade vi att “Listprice” kolumnen hade en del NULL-värden och kunde lösa det därifrån.

2. **Eget betyg**

- Svårt att säga utan referenspunkter. Tycker att jag resonerar tydligt och analytiskt i både text och kod och använt hyfsat avancerad kod för att få fram info, men att det ändå kan finnas mer värdefull data att hitta än den jag tagit fram. G/VG.

3. **Tips till mig själv**

- Att göra ett grundligt förarbete är verkligen A och O här, titta noga igenom databasen och tänk ut vad som faktiskt är värt att analysera, vad som tillför värde.
- En query behöver inte vara komplicerad för att visa värdefull data.
- Lös inte problem med utgångspunkten “här ska jag använda metoden X...”, utan använd mer avancerad kod när det faktiskt behövs.