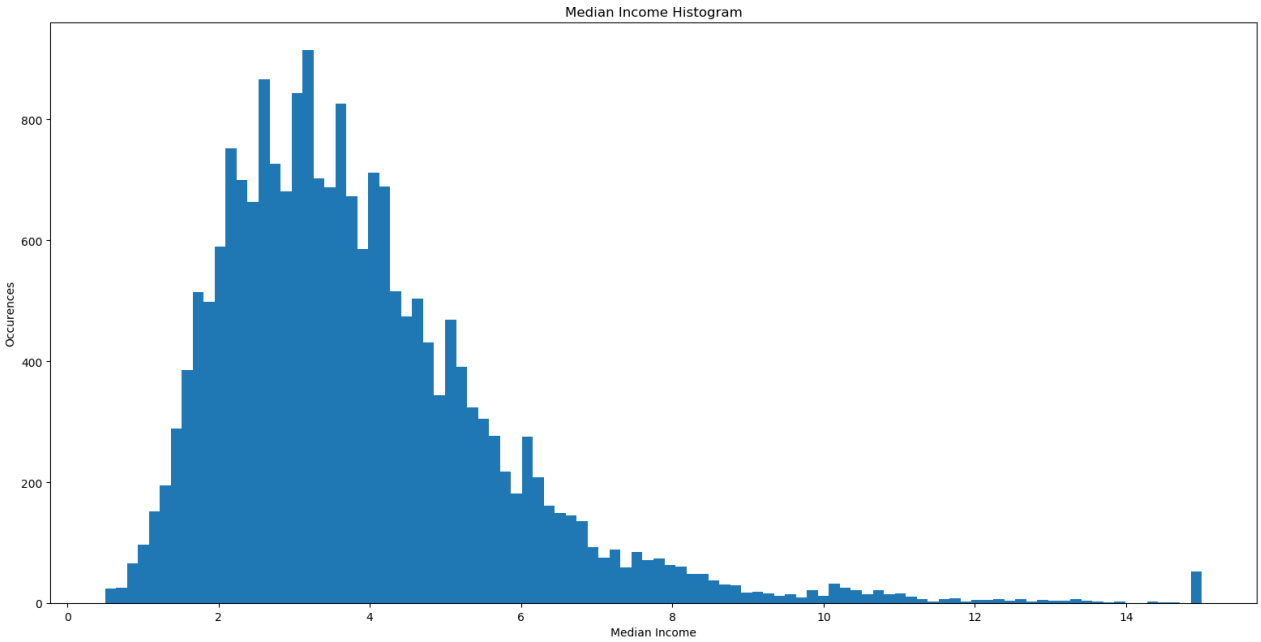
ITMAL Øvelser – Uge 5

Øvelse 1 :

a) Plot fordelingen af median_income. Find også spredning, middelværdi og median.



Figur 1 - Median Income Histogram

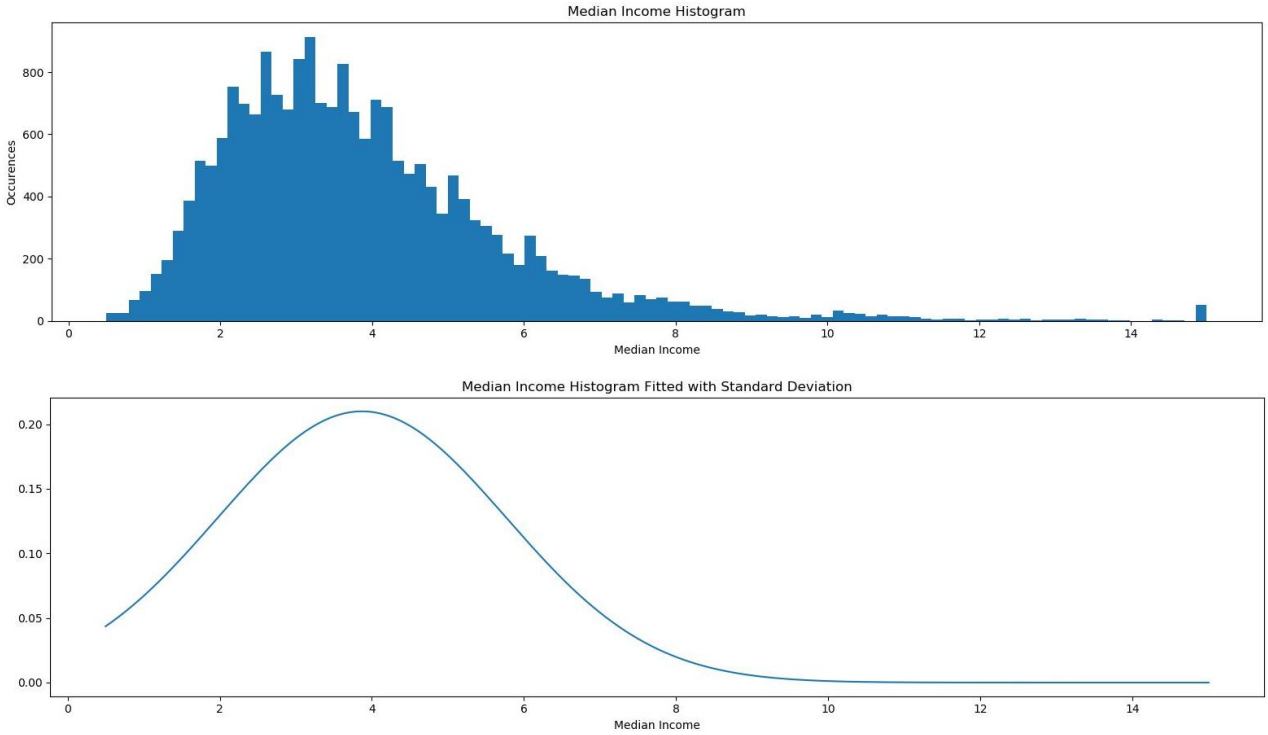
På figur 1 ses histogrammet for median_income. Dette plot viser fordelingen af median_income og hvor mange gange de forskellige værdier optræder. Nedenfor kan vi se spredningen, middelværdien og medianen:

```
Spredningen er: 1.89978
Middelværdien er: 3.87067
Medianen er: 3.53480
```

b) Er der forskel på median og middelværdi af median_income? Hvilken af de to beskriver bedst en ”almindelig families indkomst” og hvorfor?

Der er forskel på medianen og middelværdien, dette kan ses på værdierne som blev fundet i spørgsmål a. Tit er medianen det bedste billede på en almindelig families indkomst, fordi at der kan være nogle få virkelig høje værdier (dem der tjener virkelig meget!) som gør at vores middelværdi bliver skæv. For eksempel er der et ophop af høje værdier, som kan ses på median income histogrammet.

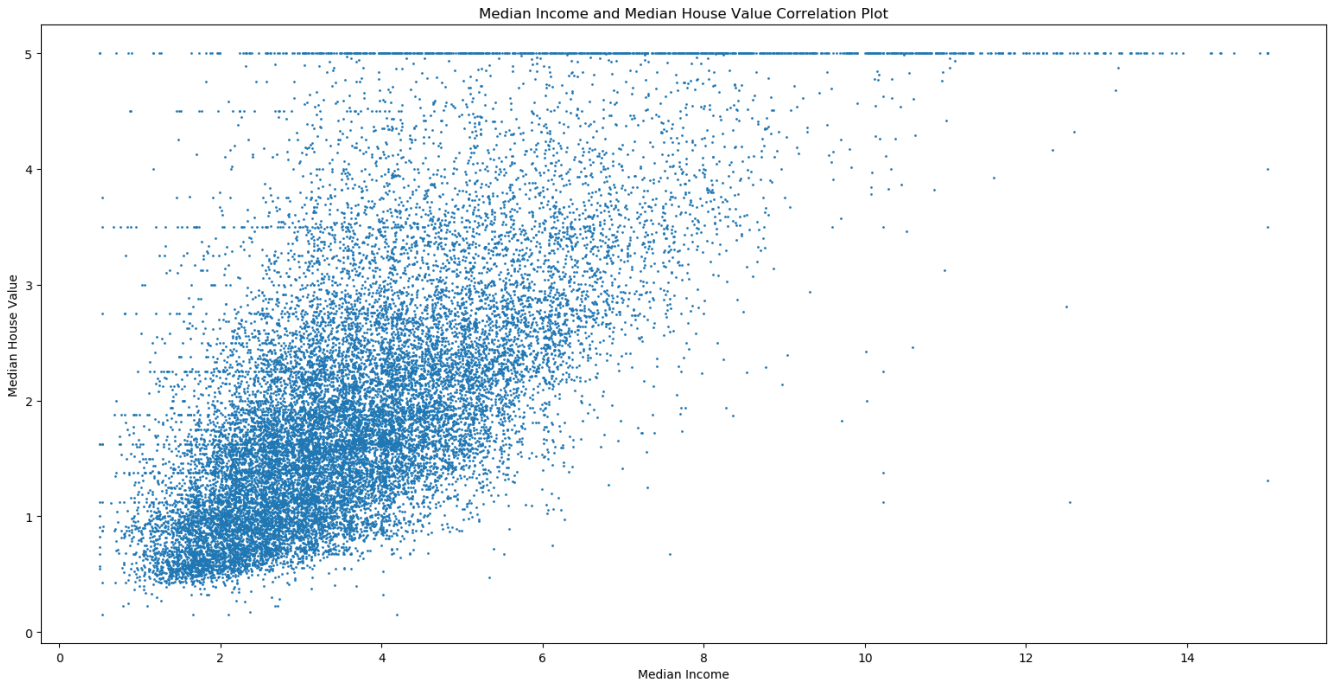
c) Fit en normalfordeling til data og plot histogrammet – passer de to?



Figur 2 - Median Income Histogram (Normalfordeling)

På figur 2 ses histogrammet for median_income, samt normalfordelingen til dataen (nederste histogram). Grafisk kan vi se at normalfordelingen på dataen passer meget godt med vores tidligere lavet data histogram.

d) Er der sammenhæng imellem median_house_value og median_income? Lav korrelationsplot.



Figur 3 - Median Income and Median House Value Correlation Plot

På figur 3 ses korrelationsplottet for median_income og median_house_value. Grafisk kan vi se at der er korrelation imellem de to variabler. Men for at få en værdi er der udregnet en værdi i Python ved hjælp af korrelation imellem de to variabler, denne udregnede værdi kan ses nedenfor:

```
Korrelations Koefficient er: 0.68808
```

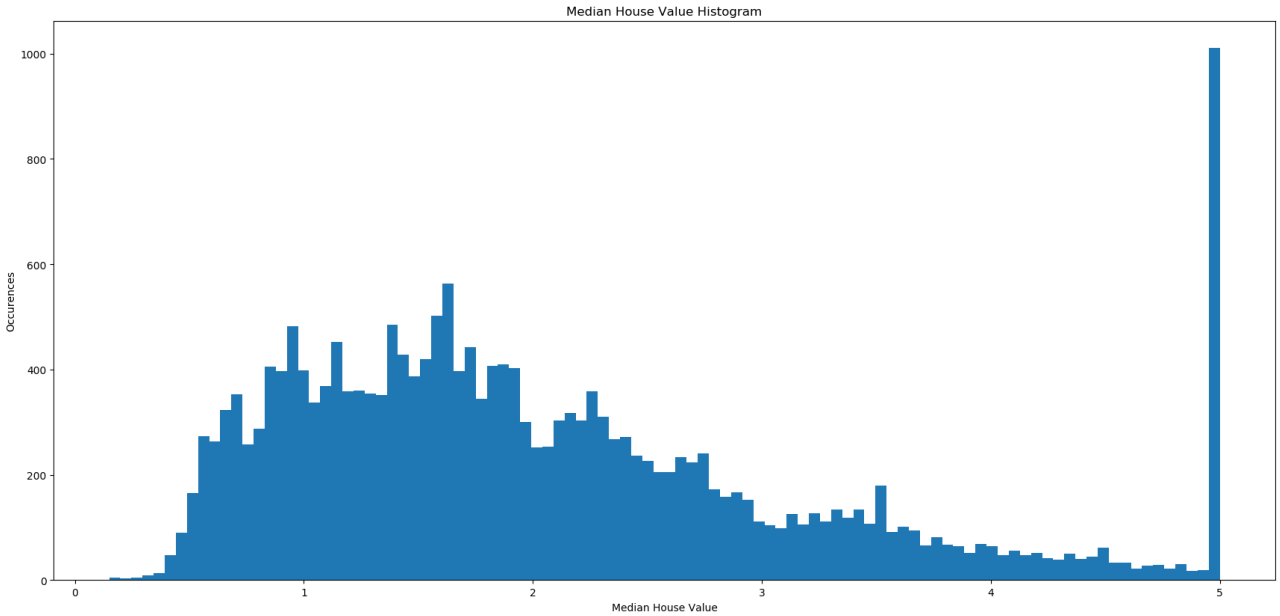
Korrelations koefficienten giver et godt billede på hvordan korrelationen er. De to variabler har ifølge koefficienten en forholdsvis høj sammenhæng, men den kunne også sagtens have været endnu højere og dermed have endnu mere sammenhæng.

e) Hvad er 5% og 95% percentilerne af median_house_value? Plot også fordelingen af median_house_value. Kommenter på realismen af max-værdi og 95% percentil – foreslå gerne en løsning til hvad man kunne gøre ved dette, hvis man skal have mere realistiske data.

Først er 5 og 95 percentilerne fundet. Disse værdier kan ses nedenfor:

```
Den 5 percentile er: 66200$
Den 95 percentile er: 489809$
```

Herudover er der lavet et histogram for de forskellige median_house_value værdier.



Figur 4 - Median House Value Histogram

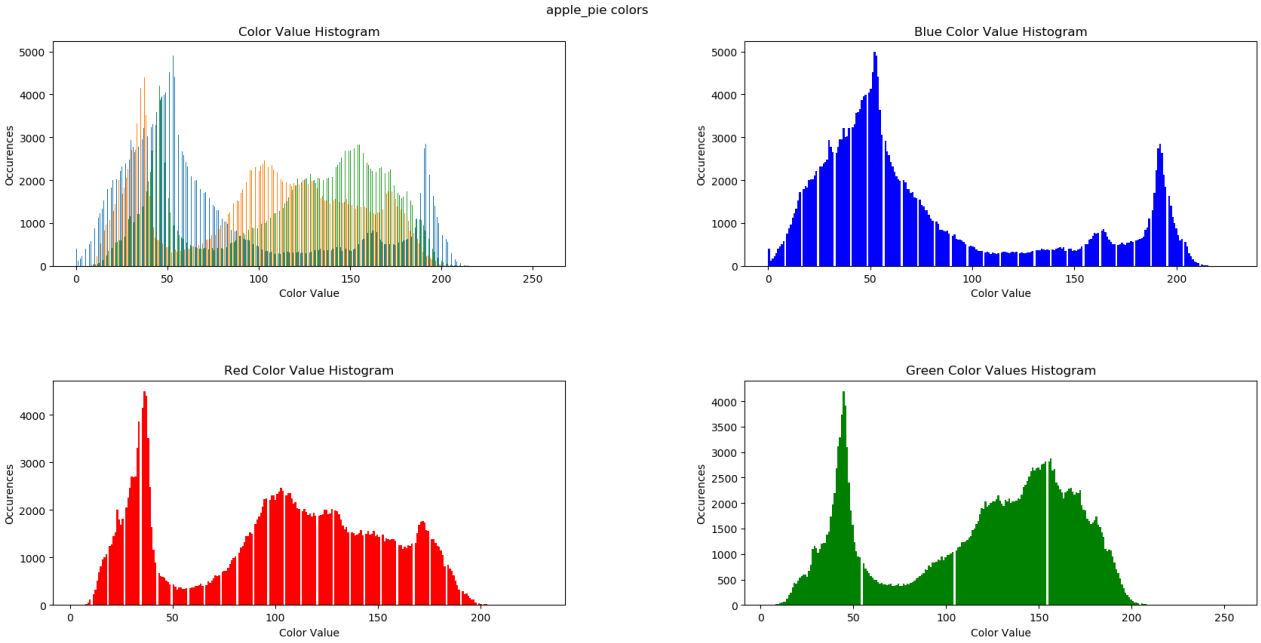
På figur 4 ses histogrammet for median_house_value. Det er tydeligt at både max-værdien og den 95 percentil bliver forskruet af et loft på max-værdien. Dette loft ser ud til at være sat på 500.000$. Dette giver et forkert billede af dataen, da hverken den 95 percentil eller max-værdi er realistiske.

En måde man kunne have løst dette kunne have været at have sat loftet højere ved indsamling af data, da vi allerede har dataen er dette selvfølgelig ikke en mulighed. En anden mulighed er at fjerne alle disse instancer, hvor median_house_value ligger ved max loftet, dette vil fjerne en masse data fra datasættet og virker derfor heller ikke som en optimal løsning. En tredje mulighed ville være at gøre loftet højere og fordele instancerne, hvor median_house_value ligger ved max loftet ud over det nye område. Her kunne man eventuelt fordele dem faldende, så der ville være færre af de virkelig høje værdi og lidt flere af værdierne som ligger lige over det oprindelige loft. Dette vil give et kvalificiret bud på, hvordan de egentlige værdier så ud uden loftet, men ville stadigvæk ikke være fuldstændige præcise. Dermed vil den tredje mulighed være den bedste løsning til at forbedre dataen.

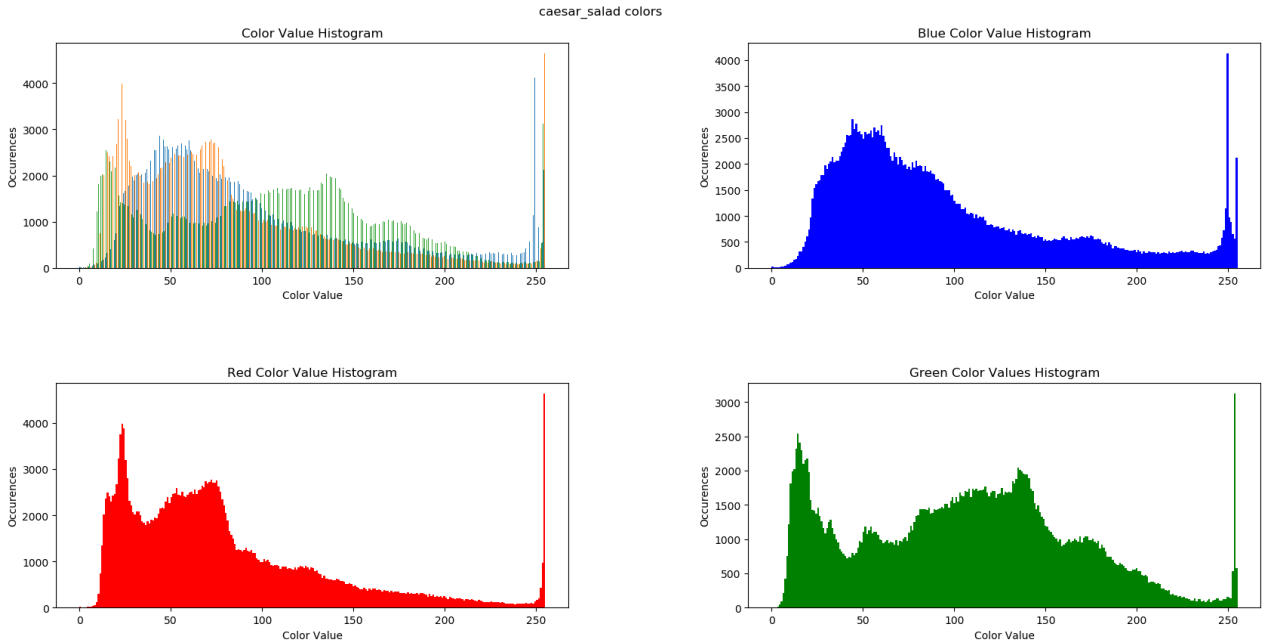
Øvelse 2 :

- Lav data analyse på jeres egne data og projekt.

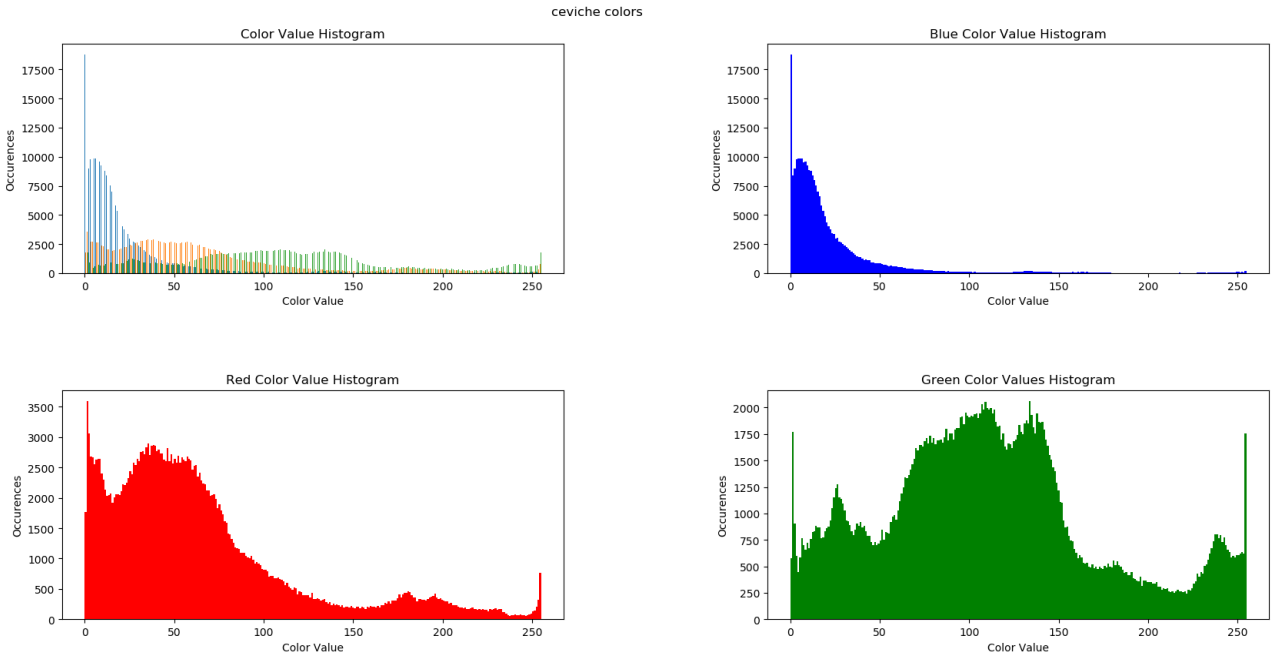
Vi har lavet analyse på vores egne data til vores projekt. Vores projekt går ud på at klassificere forskellige madretter ved hjælp af billeder af disse. Til data analyse har vi lavet histogrammer af de forskellige madretters farve bitværdier. Som udkast har vi figurer med få forskellige retter.



Figur 5 - Apple Pie Colors Histograms



Figur 6 - Caesar Salad Colors Histograms



Figur 7 - Ceviche Colors Histograms

Vi ser forskellige farve histogrammer for de 1000 forskellige billeder af Apple Pie, Caesar Salad og Ceviche. På disse histogrammer støder vi ind i forskellige problemstillinger der kan være interessante i vores projekt.

På figur 5 ser vi at der mangler forskellige værdier i vores histogram for Apple Pie af både rød, blå og grøn farve. Hvad dette specifikt betyder for vores machine learning og hvorfor de mangler er ukendt. Umiddelbart burde det ikke give store problemer i forhold til vores machine learning.

På figur 6 ser vi at der er et overtal af meget høje værdier af både rød, blå og grøn i de forskellige billeder af Caesar Salad. Dette betyder at der er et stort antal mørke farver i billederne. Her igen er spørgsmålet om dette får betydning for vores machine learning og om det er muligt at få det filtret væk ved hjælp af noget convolution.

På figur 7 ser vi at der er et overtal af meget lave værdier af både rød, blå og grøn i de forskellige billeder af Ceviche. Dette betyder at der er et stort antal lyse farver i billederne. Dette kan muligvis være fordi at ceviche er en lille ret og at der derfor er meget af tallerknerne der kan ses, hvilke vil give overflåd af hvid i billedet. Ligesom får er spørgsmålet om dette kan filtreres væk eller om det giver et godt billede af retten i stedet.