

# Maskinlæring

Våren 2018



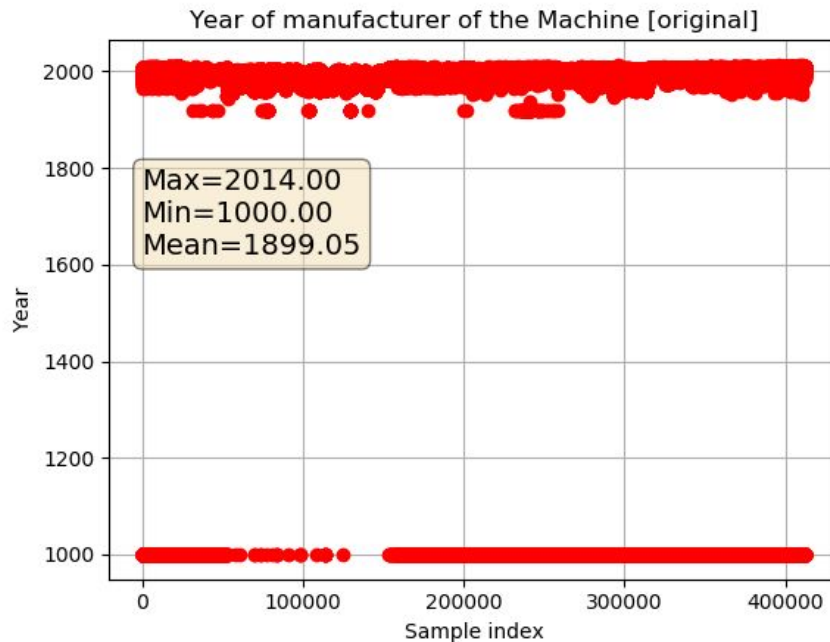
# Statistikk over utvalgte variabler

Lag en enkel statistikk for noen utvalgte variabler, både visuelt og numerisk.  
(Snitt, Max, Min).

# Statistikk over utvalgte variabler

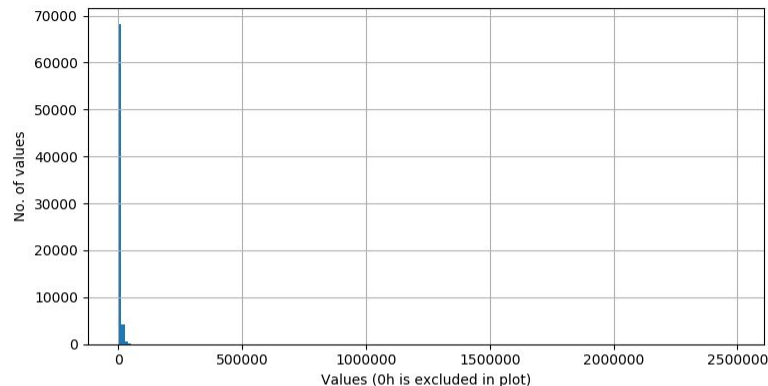
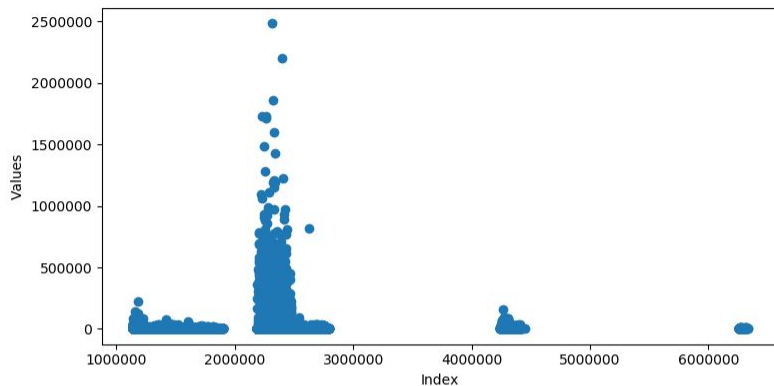
	SalesID	SalePrice	MachineHoursCurrentMeter	UsageBand	saledate	Hydraulics
count	412698	412698	147504	73670	412698	330133
unique				3	4013	12
top				Medium	2/16/2009 0:00	2 Valve
freq				35832	1932	145317
mean	2011161	31215	3523			
min	1139246	4750	0			
max	6333349	142000	2483300			

# Kikke på dataen: YearMade - Problematiske verdier

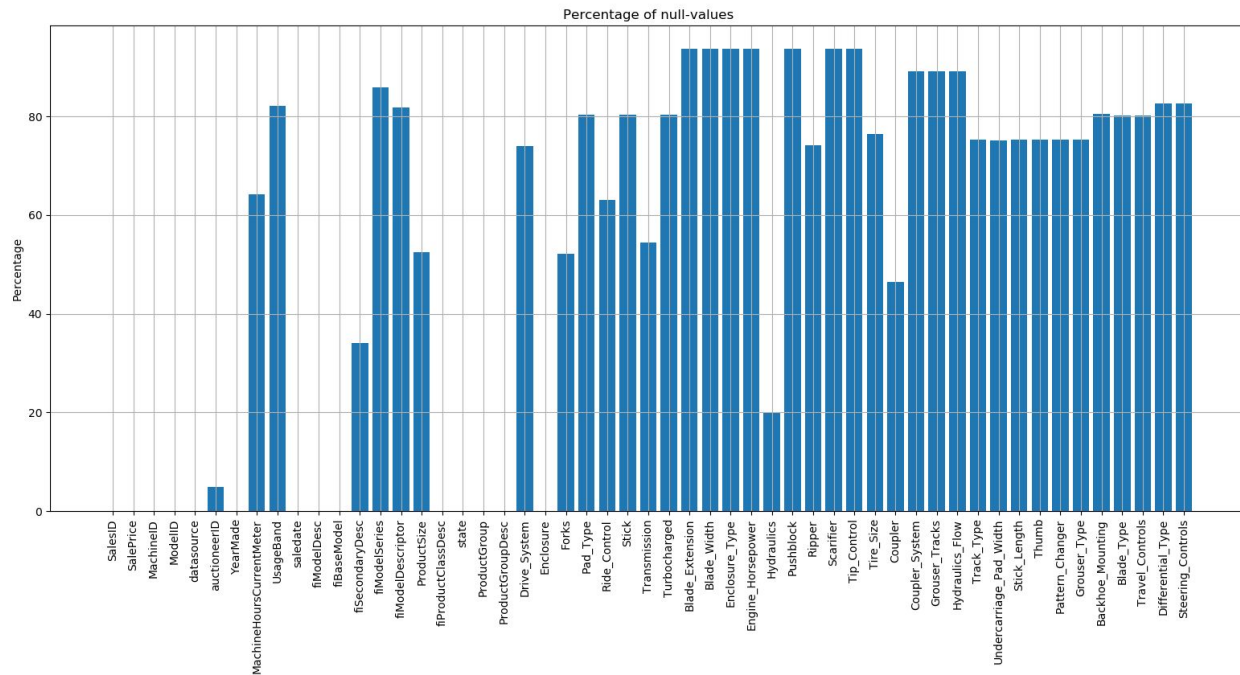


# Kikke på dataen: MachineHoursCurrentMeter - Extreme Outliers

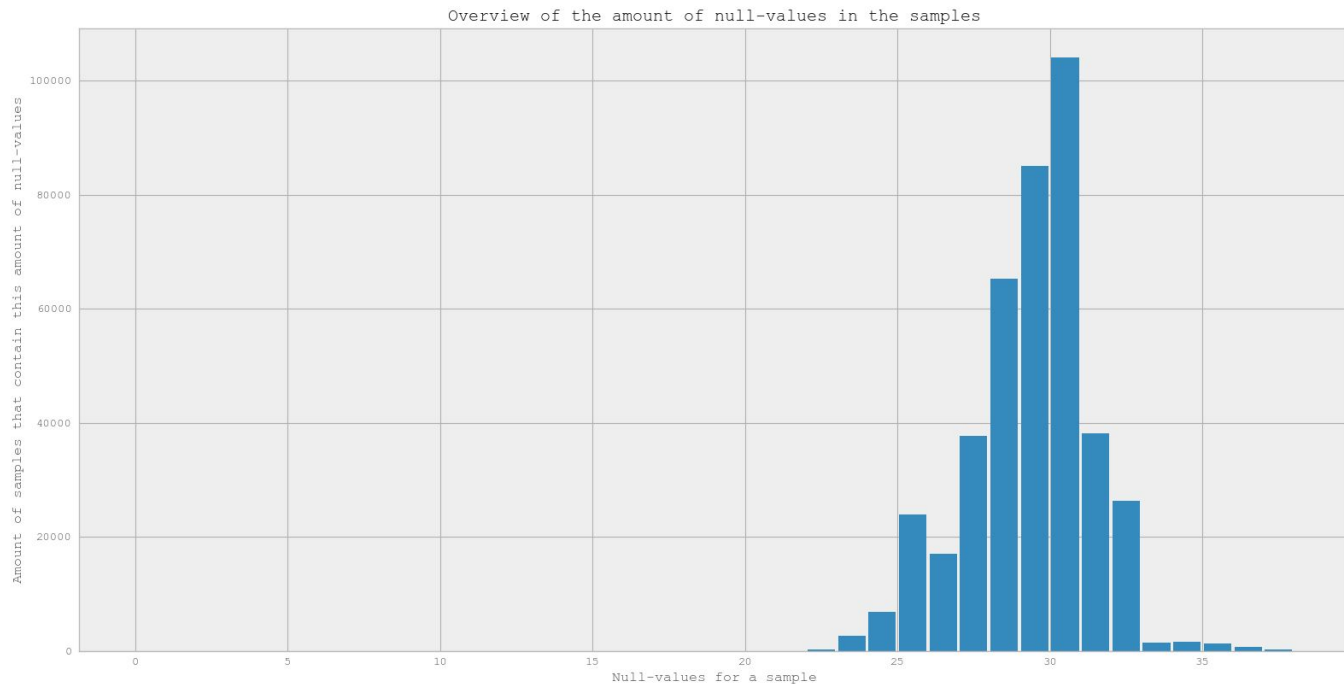
MachineHoursCurrentMeter



# Kikke på dataen: Tettheten av 'null'-verdier



# Kikke på dataen: 'Missing data' i individuelle samples



# Preprosessering: Missing data

Gjør en vurdering av mengden manglende verdier, konsekvens av dette, og nevner noen metoder for å håndtere det.

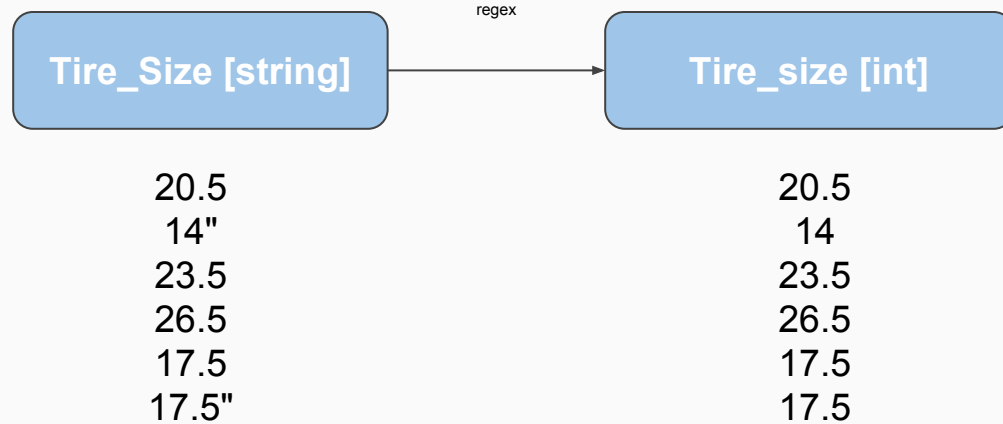


# Preprosessering: Missing data

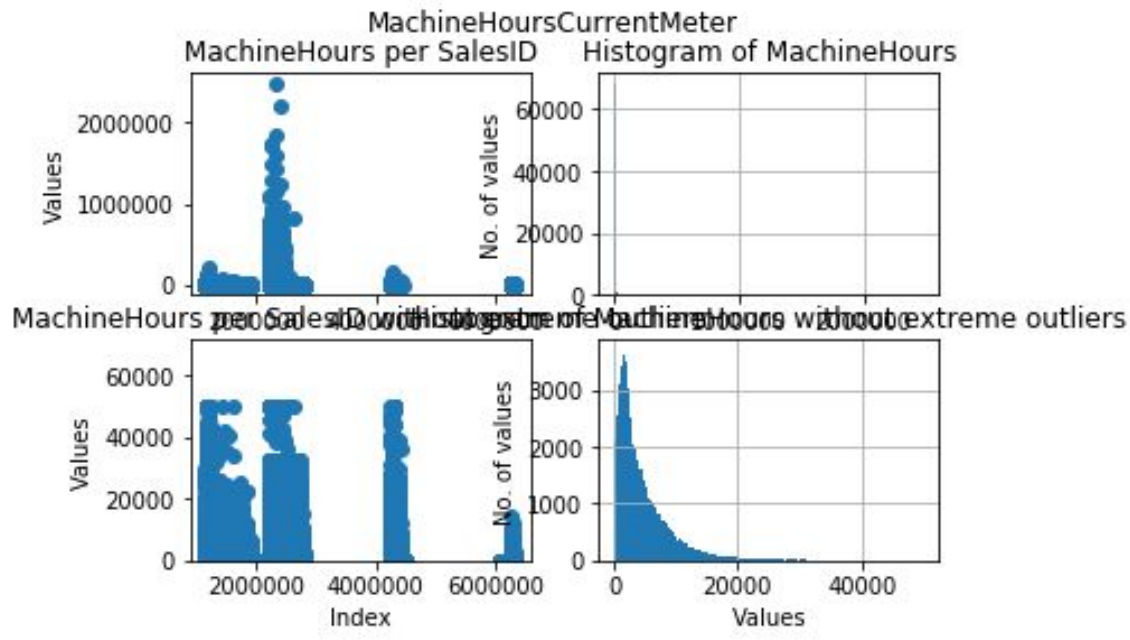
- Behandle missing data
  - Fjerne samples med missing values
  - Fjerne verdier med mye missing values
  - Estimere missing values
    - “Mean”, “median”, “most frequent” (mode)...
    - Regression Substitution
    - Multiple Imputation

# Preprosessering: Tekst til tall

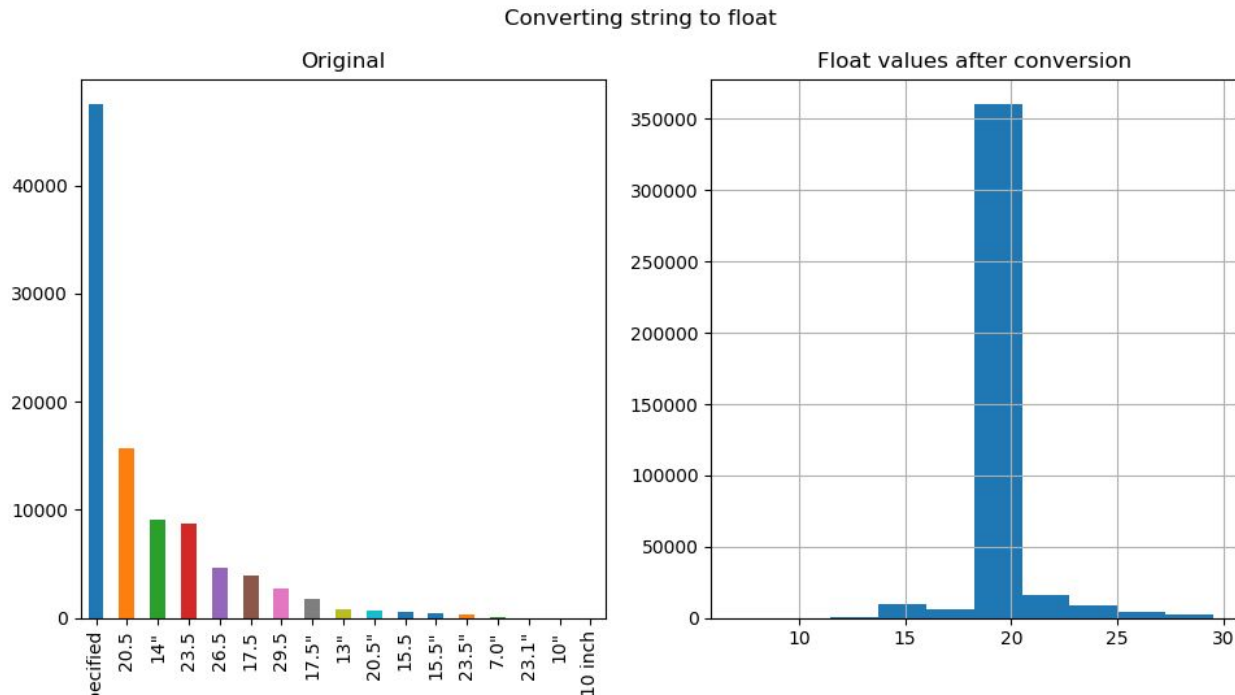
Tire\_Size  
Undercarriage\_Pad\_Width  
Stick\_Length



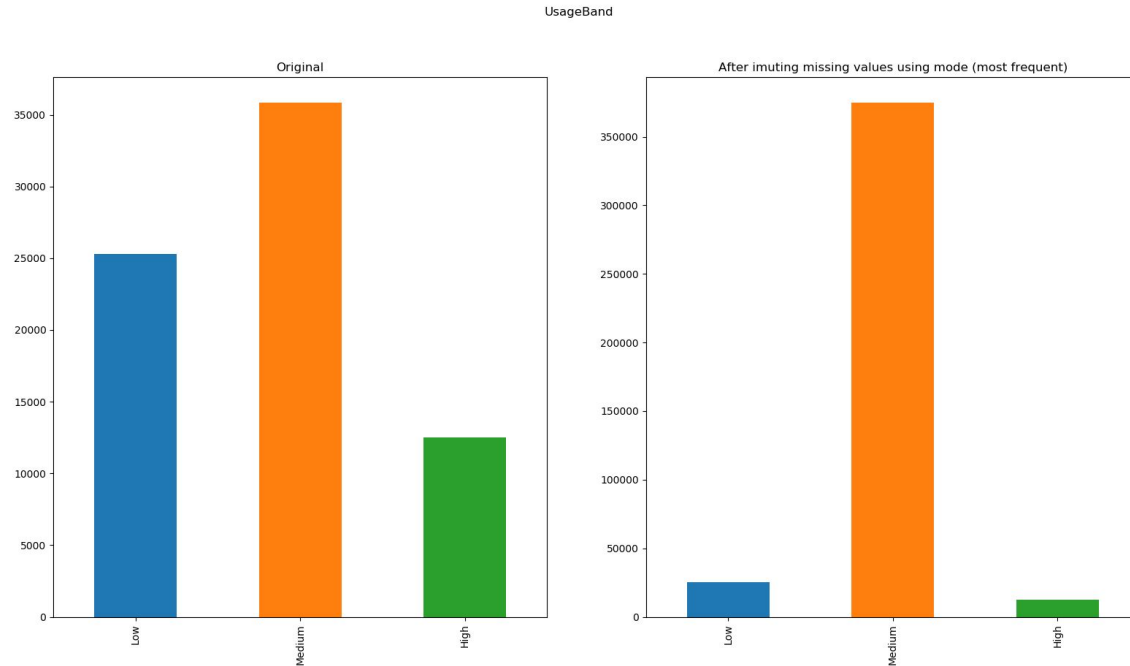
# Preprocessing MachineHoursCurrentMeter - Extreme Outliers



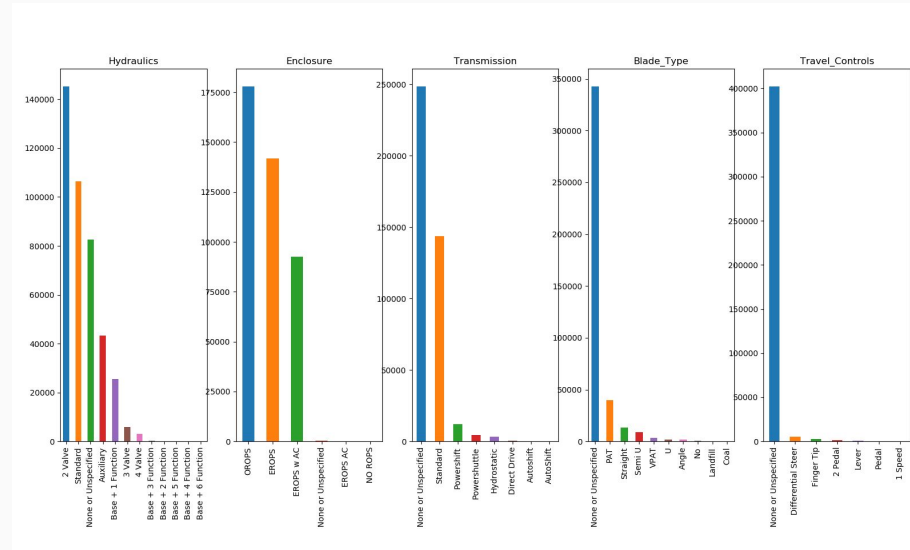
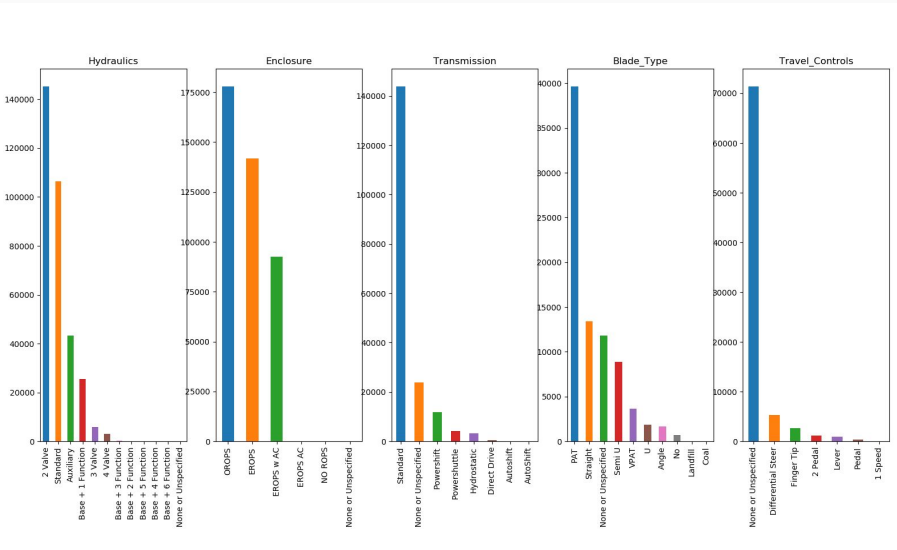
# Preprosessering: Tekst til tall



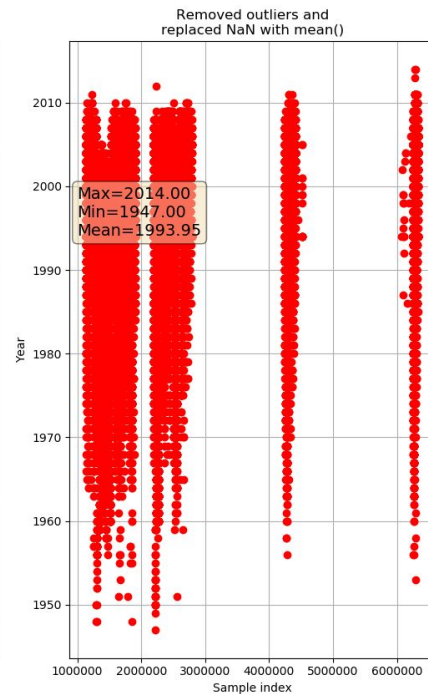
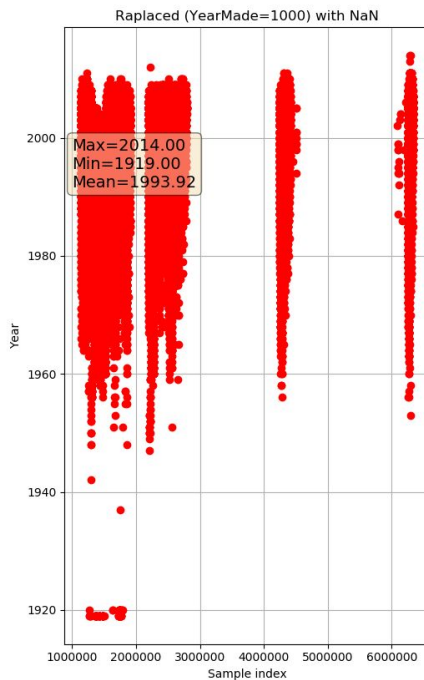
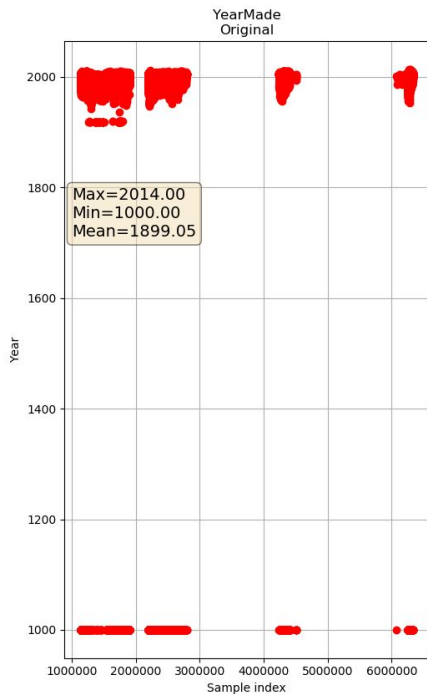
# Preprosessering: Kategorivariabler - NaN til Most Frequent



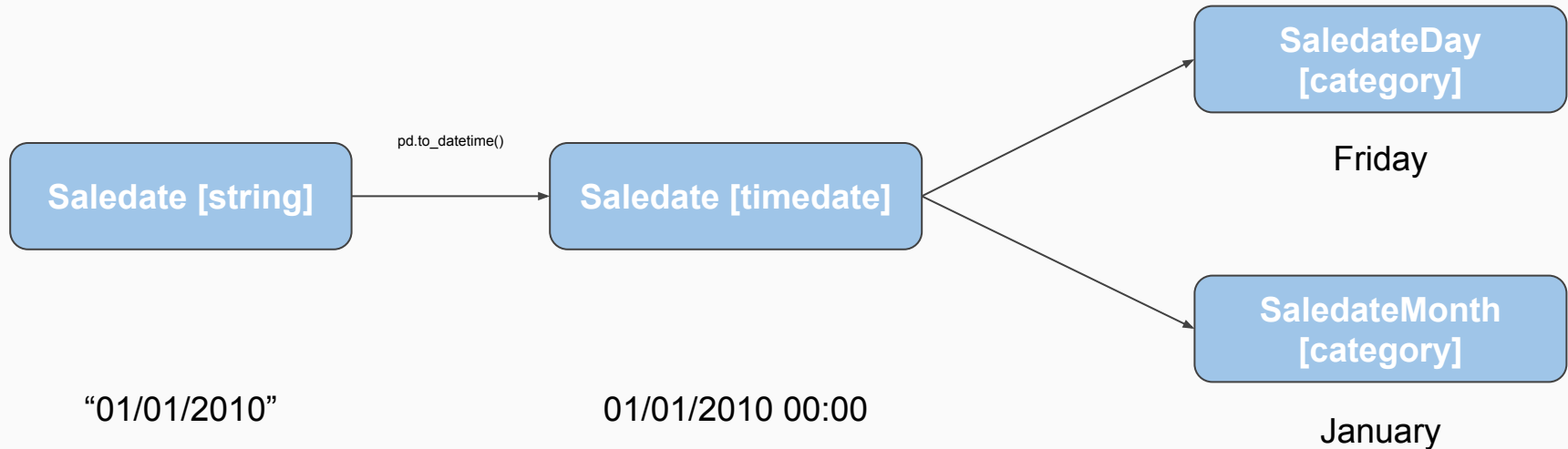
# Preprocessing: Kategorivariabler - “None or unspecified”



# Preprosessering: YearMade

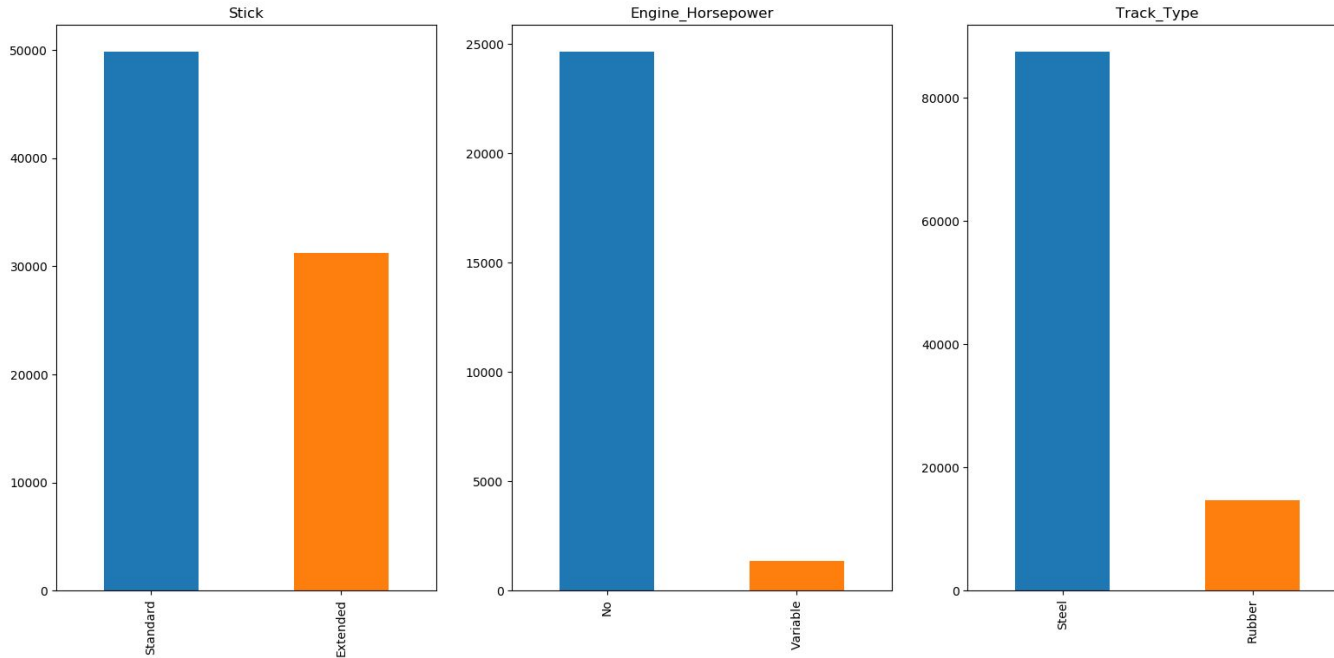


# Preprossering: Saledate - Håndtere tid





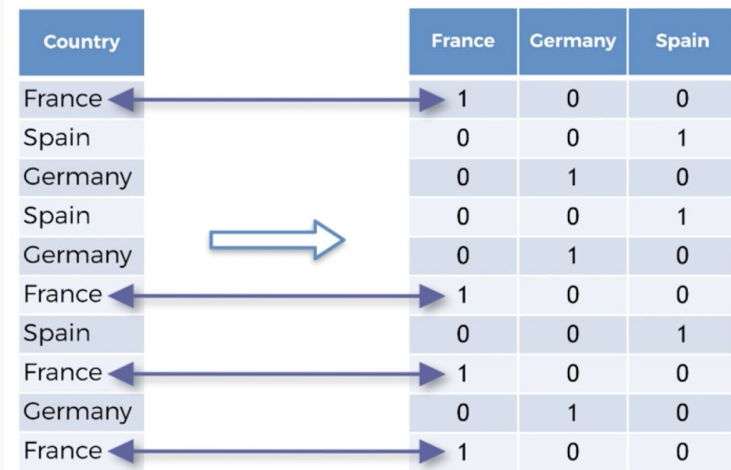
# Preprosessering: Problematiske variabler



# Klassifisering av data variabler

Gjør en vurdering om datatyper, kontinuerlige/kategoriske variable, og hvordan dette håndteres.

- ID
- Int
- Float
- Binary
- Category



Country	France	Germany	Spain
France	1	0	0
Spain	0	0	1
Germany	0	1	0
Spain	0	0	1
Germany	0	1	0
France	1	0	0
Spain	0	0	1
France	1	0	0
Germany	0	1	0
France	1	0	0

# Modellvalg

Gjør et modellvalg, og utform en kort begrunnelse som belyser fordeler og ulemper.

# Modellvalg

	<b>Ridge Regression</b>	<b>Artificial Neural Net (182 - [50 - 50] - 1)</b>
Fordeler	<ul style="list-style-type: none"><li>- God innsikt i modellen</li><li>- Bedre enn ODR på multikollinearitet</li><li>- Minimerer innvirkningen til irrelevante “features”</li></ul>	<ul style="list-style-type: none"><li>- Kan modellere ulineær oppførsel</li></ul>
Ulemper	<ul style="list-style-type: none"><li>- Fjerner ikke irrelevante “features”, slik som f.eks Lasso</li></ul>	<ul style="list-style-type: none"><li>- Black box</li><li>- Beregningsmessig tungt å lage modellen</li></ul>

# Metric

## $R^2$ score:

- Enkelt med tall som prosent, hvor 100% er best
- “Overfitting” vil gi gode en god score, selv om det ikke er ønskelig

## RMSE score:

- Enkel og beskrivende
- Verdien er relativ, brukes til sammenligning

# Resultat

Beskriv resultatene, samt hvilke features er viktigst.

# Resultat

Modell	Metric	Treningssett score	Valideringsett score
Ridge Regression	$R^2$	58.0 %	57.9 %
	RMSE	14970	15071
ANN	$R^2$	58.4 %	58.2 %
	RMSE	14907	15012

# Overfitting

Redegjør for hvordan du har tatt høyde for overfitting og hvor god overførbarhet modellen har til nye data



# Testsett vs treningssett

- Ridge Regression utilizes regularization
- 70%/30% tilfeldig trenings-/test-data

