# Datan ymmärtäminen

CRISP-DM:n toinen vaihe käsittää projektiin valitun datan ymmärtämisen. Siinä käydään läpi kaikki ne datat, jotka liiketoiminnan ymmärtämisessä on valittu projektin dataresursseiksi. Datan ymmärtämiseen lasketaan työvaiheena myös datan lataaminen työkaluun, jos datan ymmärtämiseen käytetään spesifiä työkalua. Jos dataa tulee myös useammasta lähteestä, niin vaiheessa tulee myös miettiä mahdollisten integraatioiden, eli datayhteyksien luomista ja niiden ajoittamista tiettyyn vaiheeseen.

## Alustava datan keruu

Kuvaa alle käytettävissä olevat datalähteet, niiden sijainti ja niihin yhdistämisessä käytetty metodi. Jos huomaat ongelmia datalähteeseen yhdistämisessä tai datan tuomisessa, kuvaa kyseiset ongelmat taulukon ongelma kenttään.

|  |  |  |  |
| --- | --- | --- | --- |
| Datalähde | Sijainti | Metodi | Ongelmat |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

## Datan kuvaaminen

Kuvaa datan formaatin, datan määrän (esimerkiksi kuinka monta riviä ja saraketta datan jokaisessa taulussa on) sekä määrittää onko kaikkien sarakkeiden nimet esimerkiksi löydetty. Raportissa arvioidaan myös ylätasolla, onko datan määrä esimerkiksi sellainen, että suunniteltu projekti voidaan aloittaa.

|  |
| --- |
|  |

## Datan tutkiminen

Kuvataan datan tutkimisessa tehtyjen analyysien tulokset, sisältäen ensimmäiset havainnot, mahdolliset alustavat hypoteesit, ja niiden vaikutukset jäljellä oleviin projektin osioihin. Raportin ymmärtämisen kannalta on tärkeää sisällyttää tehdyt datan visualisoinnit raporttiin, jotta raportin luettavuus ja ymmärrettävyys on parempi.

|  |
| --- |
|  |

## Datan laadun

Käytä datanlaadun tehtävässä saamiasi tuloksia, ja vertaa niitä datan ymmärtämisessä ja liiketoiminnan ymmärtämisessä asetettuihin tavoitteisiin. Kuvaa kuinka hyvin datan laatu täyttää asetetut tavoitteet, ja jos ristiriitoja löytyy, esitä keinoja niiden ratkaisemiseksi.

|  |
| --- |
|  |