

Máster MIS
Data Science 2023

Análisis de variabilidad en distribuciones Linux



Diego Monsalves Vázquez
Carlos Núñez Arenas
José Antonio Zamudio Amaya

Índice general

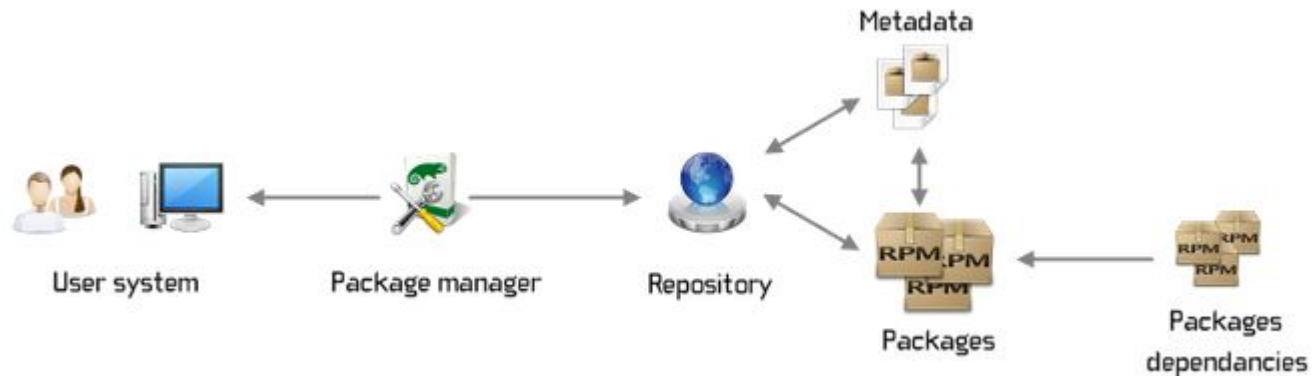
1. Dominio y objetivo
2. Preguntas de interés
3. Preprocesado
4. Análisis exploratorio
5. Resultados
6. Conclusiones

Índice general

1. Dominio y objetivo
2. Preguntas de interés
3. Preprocesado
4. Análisis exploratorio
5. Resultados
6. Conclusiones

1. Dominio y objetivo

Dominio



1. Dominio y objetivo

Objetivo

Analizar la **variabilidad** que existe en las distribuciones Linux

Analizar la **variabilidad** que existe en las versiones de una misma distribución

1. Dominio y objetivo

Utilidad

Estudiar la evolución de grandes proyectos
open-source

Obtener conclusiones del desarrollo software, y
cuestionar las ya existentes

Índice general

1. Dominio y objetivo
2. Preguntas de interés
3. Preprocesado
4. Análisis exploratorio
5. Resultados
6. Conclusiones

2. Preguntas de interés

Para test estadístico

Conforme avanza el tiempo, se añaden nuevas funcionalidades a los paquetes, o se pulen las ya existentes...

¿Cómo afecta esto al número de dependencias de un paquete respecto al tiempo?

2. Preguntas de interés

Para visualizar

En cada versión desaparecen funcionalidades
desactualizadas para dar soporte a nuevas
operaciones...

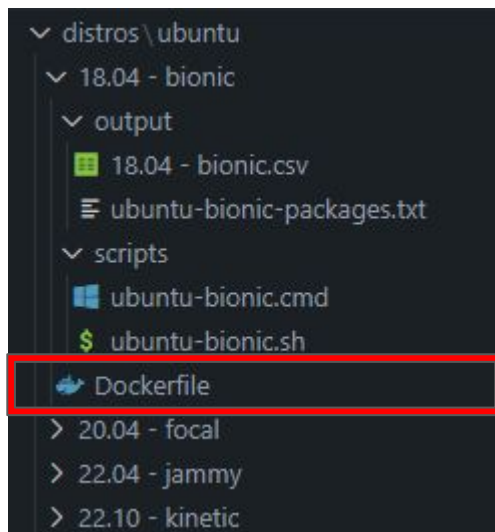
¿Cómo afecta esto al tamaño de las distribuciones?

Índice general

1. Dominio y objetivo
2. Preguntas de interés
3. Preprocesado
4. Análisis exploratorio
5. Resultados
6. Conclusiones

3. Preprocesado

Generación



Estructura

```
# Use Ubuntu Bionic as the base image
FROM ubuntu:18.04

# Update the package repository and install the necessary packages
RUN apt-get update && apt-get install -y apt-utils

# Run the command to list all packages
RUN apt-cache dumpavail | awk '/^Package:/ { package = $2 } /^Descrip

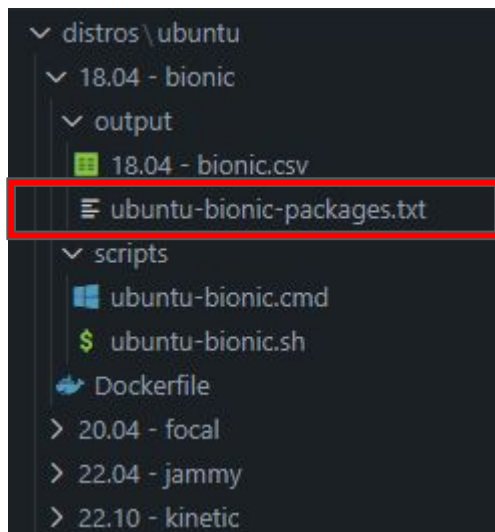
# Export the generated file outside the container
VOLUME /ubuntu-bionic-packages.txt:/ubuntu-bionic-packages.txt

# Specify the default command to run when the container starts
CMD ["cat", "/ubuntu-bionic-packages.txt"]
```

Dockerfile

3. Preprocesado

Construcción



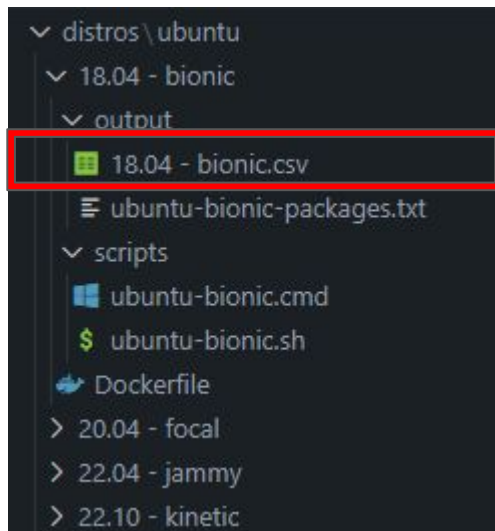
Estructura

```
Package: acct
Description: Description: GNU Accounting utilities for process and login accounting
Section: admin
Version: 6.6.4-1
Architecture: amd64
Priority: optional
Essential: no
Build-Essential: no
Important: Null
Maintainer: Ubuntu
Original-Maintainer: Debian
Size: 87216
Installed-Size: 297
Depends: libc6 (>= 2.14), lsb-base
Pre-Depends: Null
Recommends: Null
Conflicts: Null
Suggests: Null
Breaks: Null
Replaces: Null
Provides: Null
Enhances: Null
```

.txt

3. Preprocesado

Transformación



Package	Description	Section	Version	Architecture	Priority	Essential	Build-Essential
acct	Description	admin	6.6.4-1	amd64	optional	no	no
acl	Description	utils	2.2.52-3build1	amd64	optional	no	no
acpi-support	Description	admin	0.14	amd64	optional	no	no
acpid	Description	admin	1:2.0.28-1ubuntu1	amd64	optional	no	no
adduser	Description	admin	3.116ubuntu1	all	important	no	yes
adium-theme-ubuntu	Description	gnome	0.3.4-0ubuntu4	all	extra	no	yes
adwaita-icon-theme	Description	gnome	3.28.0-1ubuntu1	all	optional	no	yes
aisleriot	Description	games	1:3.22.5-1	amd64	optional	no	yes
alembic	Description	python	0.9.3-2ubuntu1	all	optional	no	yes
alsa-base	Description	sound	1.0.25+dfsg-0ubuntu1	all	optional	no	yes
alsa-utils	Description	sound	1.1.3-1ubuntu1	amd64	optional	no	yes
anacron	Description	admin	2.3-24	amd64	optional	no	yes
apg	Description	admin	2.2.3.dfsg.1-5	amd64	optional	no	yes
app-install-data-gnome	Description	gnome	15.1	all	optional	no	yes
app-install-data-qt5	Description	x11	16.04	all	optional	no	yes
apport-symptoms	Description	utils	0.2	all	optional	no	yes
appstream-glib	Description	doc	0.7.7-2	all	optional	no	yes
apt-clone	Description	admin	0.4.1ubuntu2	all	extra	no	yes

Estructura

.CSV

3. Preprocesado

Limpieza

```
# Eliminar filas duplicadas
df = df.drop_duplicates()

# Cambiar Null por NaN
df = df.replace("Null", np.nan)

# Nos quedamos sólo con las filas que tengan algo de información usable, para
# calculamos la cantidad de valores no nulos por fila
row_counts = df.count(axis=1)
# Define el umbral de cantidad mínima de valores no nulos que debe tener cada fila
threshold = 5
# Filtra las filas que tienen menos valores no nulos que el umbral definido
df = df[row_counts >= threshold]
```

Limpieza mínima

3. Preprocesado

Datos sobre el dataset

Columnas: 23

Filas: 299.733

Tamaño: 170 MB

Tiempo en generarse: +12h (optimizado a 24m)

Índice general

1. Dominio y objetivo
2. Preguntas de interés
3. Preprocesado
4. Análisis exploratorio
5. Resultados
6. Conclusiones

4. Análisis exploratorio

XX

4. Análisis exploratorio

XY

4. Análisis exploratorio

XZ

Índice general

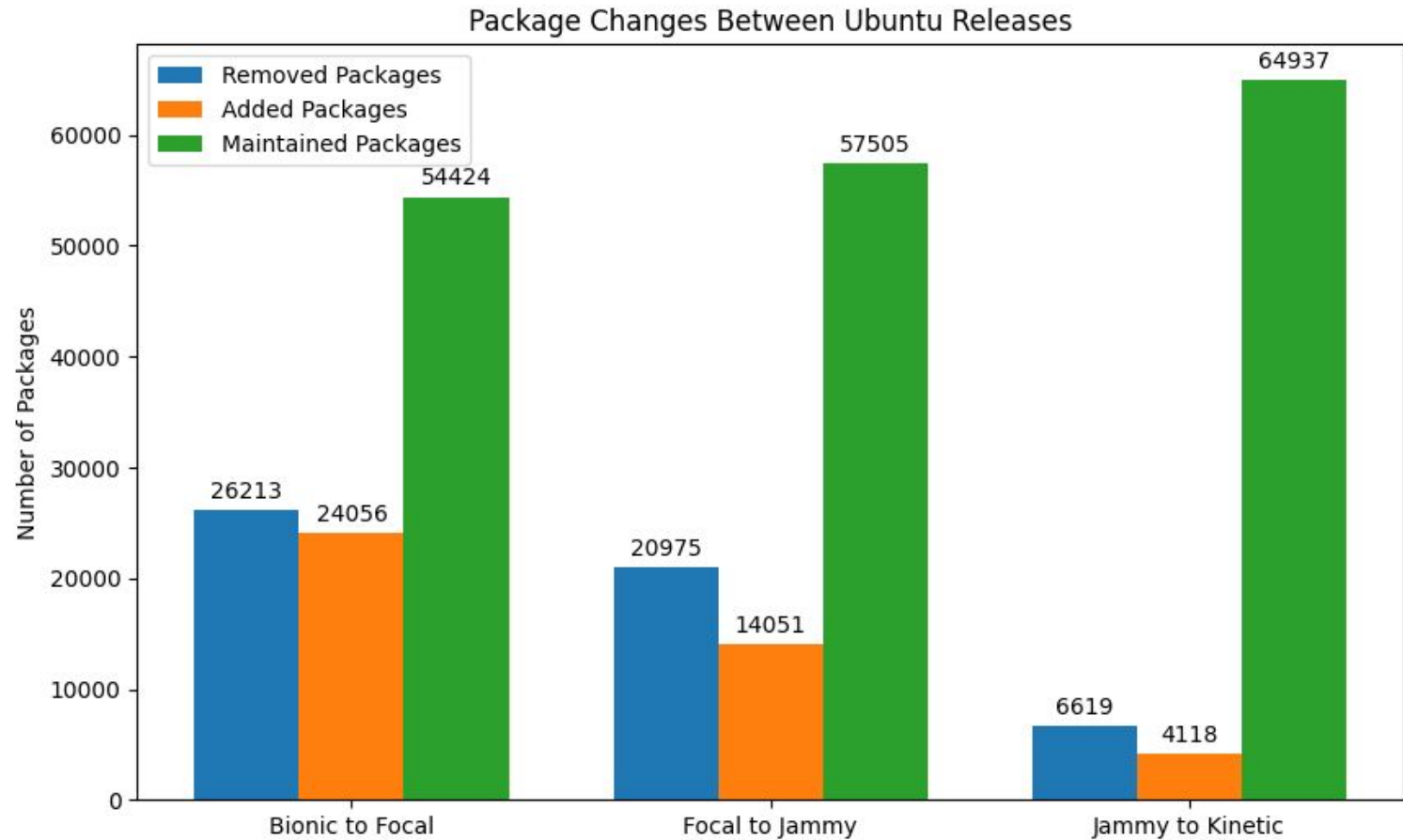
1. Dominio y objetivo
2. Preguntas de interés
3. Preprocesado
4. Análisis exploratorio
5. Resultados
6. Conclusiones

5. Resultados

Test estadístico

5. Resultados

Visualización



Índice general

1. Dominio y objetivo
2. Preguntas de interés
3. Preprocesado
4. Análisis exploratorio
5. Resultados
6. Conclusiones

5. Conclusiones

Lecciones aprendidas

- Coste de generar datasets propios
- Importancia de preprocesar la información
- La visualización no es suficiente
- Ajustarse bien a la evaluación

5. Conclusiones

Trabajo futuro

- Aumentar el número de versiones para Ubuntu
- Añadir nuevas distribuciones
- Realizar un análisis global
- Redactar artículo

Máster MIS
Data Science 2023

Análisis de variabilidad en distribuciones Linux



Diego Monsalves Vázquez
Carlos Núñez Arenas
José Antonio Zamudio Amaya